A Systematic Review of 'Fair' AI Model Development for Image Classification and Prediction

A Systematic Review of 'Fair' AI Model Development for Image Classification and Prediction

Cite this Accepted Manuscript (AM) as Accepted Manuscript (AM) version ofRamon Correat, Mahtab Shaan, Hari Trivedi, Bhavik Patel, LeoAnthonyG. Celi, JudyW. Gichoya, Imon Banerjee, A Systematic Review of 'Fair' AI Model Development for Image Classification and Prediction, Journal of Medical and Biological Engineering https://doi.org/10.1007/s40846-022-00754-z

A Systematic review of 'Fair' AI model development for image classification and prediction

Ramon Correa[1], Mahtab Shaan[2], Hari Trivedi[3], Bhavik Patel[2,1], Leo Anthony G. Celi[4,5,6], Judy W. Gichoya[3], and Imon Banerjee[2,1,*]

[1]School of Computing and Augmented Intelligence, Arizona State University, USA
[2]Department of Radiology, Mayo Clinic, Arizona, USA
[3]Department of Radiology, Emory University School of Medicine, Atlanta, USA
[4]Massachusetts Institute of Technology, Institute for Medical Engineering and Science, Cambridge, MA
[5]Harvard TH Chan School of Public Health, Department of Biostatistics, Boston, MA
[6]Beth Israel Deaconess Medical Center, Department of Medicine, Boston, MA, United States of America
[*]corresponding author: Banerjee.Imon@mayo.edu

October 5, 2022

Abstract

Purpose - The new challenge in Artificial Intelligence (AI) is to understand the limitations of models to reduce potential harm. Particularly, unknown disparities based on demographic factors could encrypt currently existing inequalities worsening patient care for some groups.
Method - Following PRISMA guidelines, we present a systematic review of 'fair' deep learning modeling techniques for natural and medical image applications which were published between year 2011 to 2021. Our search used Covidence review management software and incorporates articles from PubMed, IEEE, and ACM search engines and three reviewers independently review the manuscripts.
Results - Inter-rater agreement was 0.89 and conflicts were resolved by obtaining consensus between three reviewers. Our search initially retrieved 692 studies but after careful screening, our review included 22

manuscripts that carried four prevailing themes; 'fair' training dataset generation (4/22), representation learning (10/22), model disparity across institutions (5/22) and model fairness with respect to patient demographics (3/22). We benchmark the current literature regarding fairness in AI-based image analysis and highlighted the existing challenges. We observe that often discussion regarding fairness are limited to analyzing existing bias without further establishing methodologies to overcome model disparities.
Conclusion - Based on the current research trends, exploration of adversarial learning for demographic/camera/institution agnostic models is an important direction to minimize disparity gaps for imaging. Privacy preserving approaches also present encouraging performance for both natural and medical image domain.

Keywords: Fairness in AI, Medical Imaging, Natural Imaging, Image Classification.

1 Introduction

The development of Artificial intelligence (AI) models, particularly deep learning, in medical imaging has achieved remarkable performance such that models are able to match expert-level accuracy. For example, performance at par with specialists has been shown in a variety of clinical applications, from diagnosis of common thoracic pathologies [1] to detecting diabetic retinopathy on retinal fundoscopic images [2]. Interestingly, some recent studies have claimed that AI has surpassed expert-level performance. For example, [3] showed that AI had a higher sensitivity for detecting abnormalities on screening mammograms when compared to a radiologist, particularly for subtle lesions, [4] showed that analysis of brain MRI using AI has greater sensitivity in identifying parenchymal changes reflective of early ischaemic stroke within a narrower time window from onset of symptoms with greater sensitivity than a human reader.

However, in parallel with this remarkable success, recent works have raised concerns towards the risk of unintended bias in AI systems affecting individuals unfairly based on race, gender, and other clinical characteristics [5, 6, 7]. In statistics, bias is defined as an error generated from erroneous assumptions used in the learning algorithms. There are two types of bias that can affect AI models. One is 'algorithmic AI bias' where algorithms are trained using biased data while 'societal AI bias' is caused by the assumptions and norms imposed by the society. For example, AIs trained on news articles show a bias against women. Societal bias often significantly influences algorithmic AI bias. Examples include differing allocation

of healthcare resources based on patient demographics [8, 9], bias in language models developed on clinical notes [10], and melanoma detection models developed primarily on images of light-colored skin [11]. Beyond healthcare applications, similar biases have been observed for natural images, such as face detection where models fail to correctly identify individuals of minority groups [12].

A core challenge for analyzing AI bias is that the reasons resulting in unfair AI models are not mutually exclusive and can often exacerbate one another. Recently, [13] showed that AI models trained for diagnosis can learn unintended racial information from different imaging modalities. Thus, AI models may use learned demographic information for detecting a diagnosis even when such attribute is not associated with the diagnosis. There are examples of race-ethnicity and gender influencing clinician decision-making [14], and given that AI is trained on real-world data, it is not far-fetch to assume that computers will learn to do the same.

Medical imaging datasets used to train and validate algorithms tend to originate from single health systems where the patient demographics may only be representative of a segment of the population creating data deserts without representation in AI training data [15]. Moreover, clinical practice patterns, which influence the relationship between patient and disease features with diagnosis and clinical outcomes, vary across health systems and over time. Local practice pattern changes and other extraneous factors, such as adoption of new clinical protocols, are not usually reflected in the EMR. This often brings into question how reliable models generalize to populations across space and time, how these models might fail, and whether they will entrench or even scale health disparities. Previous works have only sought to study the generalization of model performance across institutions without consideration of demographic shifts across institutions or over time [16]. There is a critical need to evaluate how individual, institutional and societal biases can be mitigated or exacerbated by adopting AI models.

To our knowledge, no systematic review has aggregated information regarding imaging-based AI models contributing to disparities and methods of systematically reducing model bias. We believe that it is a major rate limiting step for developing 'fair' AI model. This review will identify existing best practices in developing AI models for both natural and clinical images as well as gaps in the evaluation of and mitigation of the impact of algorithms on outcome inequities. In addition, we also highlight the challenges on how the best practices will be implemented and how the gaps will be addressed.

1.1 Terminologies

We briefly define the terms and their acronyms that have been referenced throughout the paper as well as in the 'fair' AI field.

- *'Fair' AI model.* Ideally, a 'Fair' AI model is defined as a model without any disparate treatment and disparate impact to certain unpledged groups based on protected attributes like gender, race, religion, color, age, and more.
- *Model optimization.* In brief, a machine learning model learns to map a set of inputs to a set of outputs from training data using the best weights. Learning which weights are the best for learning, is achieved through search optimization by navigating a set of possible weights towards the most optimal task performance.
- *Performance measures, AUC, mAP, MAE.* Model evaluation aims to estimate the generalization accuracy of a model on unseen test data. Receiver operating characteristic curve (ROC) plots true positive rate vs. false positive rate at different classification thresholds. Area under the curve (AUC) measures the entire two-dimensional area underneath the ROC curve and is commonly measured for binary classification problems. The mean Average Precision (mAP) score is calculated by taking the average precision over all classes. Mean absolute error (MAE) is a measure of errors between the true and model predicted values, and is primarily used for regression task.
- *Model Disparity.* Disparity can be defined as difference in model performance across different population subgroups. Inadequate or bias training data may lead to suboptimal AI models with high performance for majority of training data but inherent bias for minority groups, which may have profound negative impacts on health care.
- *Disparity measures.* Common metrics to measure the model disparities include difference in true positive rate (TPR disparity) and false negative rate (FNR disparity) across different subgroups against the reference group (majority class).

2 Method

We conducted a systematic review of 'fair' AI model development in imaging following the PRISMA guidelines [17]. A search for natural images

included images from landscape, person, as well as aerial, cartoon, and LiDAR images. Medical imaging search included all studies of images related to various body-parts acquired for diagnostic and treatment purposes, e.g: X-ray, H&E slides, magnetic resonance images. The following subsections respectively describes our search criteria, results, and summary.

2.1 Search strategy

Our search was obtained by querying PubMed, IEEE's, and ACM's engines and reviewed by the team in Covidence. The results are aggregated together and screened as specified in Fig. 1. The search query was "(disparity OR bias OR fairness) AND (classification OR prediction) AND (image) AND (deep learning[MeSH Terms])" present within the abstract of the article. The queries were limited to the last ten years of publication (Jan, 2011 - July, 2021) and the systems developed to support image classification and prediction tasks. Models for image segmentation and object detection task were excluded due to distinct methodological challenges for bias removal.

2.2 Study selection

Our study screening process involved *three independent reviewers* who screened a total of 692 studies retrieved by the initial search in Covidence. Inter-reader agreement was assessed using Fleiss' kappa score. Papers were excluded if they met one of the following criteria i) no peer review (e.g. published in arXiv, medRxiv, institutional databases); ii) the main topic was dataset curation and no model development and training was proposed; (ii) not a deep learning model. 71 papers were assessed for eligibility and, finally, included the papers which either discuss the disproportionate performance of models and/or propose novel techniques for mitigating model bias.

3 Results

Figure 1 presents the PRISM selection flowchart. Inter-rater agreement was 0.89 and conflicts were resolved by obtaining consensus between three reviewers. Based on the consensus of all three reviewers, 22 papers satisfied our inclusion criteria and were selected for final extraction and review. A summary of the selected literature is presented in 3 tables - studies applied to natural images: Table 1, 2, and medical images: Table 3. Figure 2 shows the diagrammatic representation of the study grouping strategy for this review. We adopted five benchmarking criteria to analyze the papers - (1)

Dataset origin: training and evaluation data, (2) *Modality*: targeted imaging modality, (3) *Method for de-biasing*: strategy for mitigating bias, (4) *Targeted task*: actual prediction or classification objective, and (5) *Performance*: final model performance reported (after de-biasing if novel mitigation strategy proposed). We present a summary of the papers in terms of model fairness in the *Observation* column.

3.1 Natural imaging study grouping

We grouped the debiasing methods proposed for natural imaging into two core themes - (1) *intelligent data sampling to reduce implicit training data bias* and (2) *fair representation learning for better understanding of data.*

3.1.1 'Intelligent' data sampling

Often bias introduced in the model due to certain elements which are more heavily represented than others within the training samples, and thus resulted skewed outcomes, low accuracy levels. [18] performed an interesting experiment *'Name that dataset'* and showed that simple model can correctly identify the source of LiDAR images which is feasible since datasets appear to have strong intrinsic bias. Authors proposed fine-tuning strategies to compensate for such dataset biases without modifying the model training scheme. While fine-tuning is often limited by accessibility of the external dataset, [19] developed a crowdsourcing data sampling framework for obtaining a 'fair' training dataset for face images. The sampling leverages an efficient batch-level demographic label inference model and a joint accuracy-aware data shuffling method. By increasing the demographic diversity, authors were able to train a model which achieve significant performance improvements - actual positive rate parity decrease from 0.282 to 0.055 when comparing a randomly sampled dataset.

[20] identified that demographics (e.g. gender, race) introduced separate apparent biases for predicting person's age and they proposed target task adjustment approach. They clustered individuals into known groups to obtain an apparent age estimate and by training models to predict apparent age followed by a constant adjustment term for each group, the authors reduced the 10% mean absolute error on their test dataset. Similarly, [21] present a hierarchical approach that combines outputs from the Microsoft Emotion API algorithm with a specialized learner that used anthropometric features, to reduce the bias of facial emotion recognition for ambiguous categories like fear and surprise. They show that this methodology can increase

overall recognition results by 17.3%.

3.1.2 Representation Learning.

Representation learning mainly focuses on automatically learning relevant representations of the data (typically as vectors/embeddings) by extracting useful information for building classifiers or other predictors [22]. However, multiple previous works [23, 24] have shown that deep learning models could achieve good performance by learning confounding features from sensitive attributes but such models are particularly bias towards the minority classes.

To learn a fair representation of the training data, [23] proposed privacy-preserving technique as a methodology for facial image classification. They developed SensitiveNets which generates a learned embedding space that eliminates sensitive information by using an adversarial regularizer and triplet loss. The model's effectiveness was compared against a traditional model; overall accuracy decreased by 5% while adversarial tasks prediction accuracy (gender and ethnicity) dropped by 81% and 66%, respectively. [25] introduced a minimax adversarial framework that aims to learn a new representation of the facial images which maximize informativeness of the target features while being minimally informative to predict the sensitive features. They observe a target AUC of 0.97 in the CelebA dataset while sensitive attribute Gender achieved close to random prediction (AUC 0.51). [26] created an encoder/decoder architecture for enforcing fairness in 'representation' learning where a mapping will be learn from an input domain to a 'fair' target domain. Unlike other fair representation learning methods that learn separate latent embeddings, the method learns a representation in the same space as the original input data. When comparing recognition ability in CelebA dataset overall accuracy decreased, yet achieved an increased TPR for the male demographic. [27] added demographic attributes in an attribute-aware loss function that contains an additional term for an individual's age and gender, to increase performance in facial recognition task. Comparison of the feature representations showed improved class separation among various demographics resulting in improved identification quality across different gender and age groups. [28] trained a model with adversarial loss and showed a vanished correlation between the bias and the learned features on a synthetic dataset, a medical images (containing task bias), and a dataset for gender classification (containing dataset bias).

[24] developed a 'fair' camera-view agnostic model for person identification (re-ID) using a view classifier branch preceded by a gradient reversal layer where the reversal layer learns a 'representation' which is not influenced

by the camera view. An ablation study that swapped DukeMTMC [29] and Market1501 [30] datasets for training and test sets revealed variable performance yet positive gains. Similarly, [31] developed an camera-view agnostic re-ID model but they designed an unsupervised asymmetric distance learning approach for cross-view clustering. When evaluated on large re-ID datasets (e.g. VIPeR [32], Market-1501 [30]), the model achieved higher performance than state-of-the-art models.

To reduce the cross-domain representation bias, [33] developed a aerial scene classification by using distance metric learning where a cross-domain color features and bag of convolution features were used to force samples belonging to the same class to be more "similar" regardless of domain. They reported a decreases in class variability at inference and improvement in cross-domain performance. [34] developed a depth prediction model to generalize onto new domains by leveraging disparity measurements alongside with confidence estimators by a novel confidence-guided loss function that handles the measured disparities as noisy labels weighted according to the estimated confidence. Weighting model losses across domains resulted in a fair model whose mean absolute error on the new domain decreased consistently. [35] presented a low-rank parameterized CNN model for end-to-end domain generalization learning and outperformed benchmark methods for a more complex mix domains of photo, sketch, cartoon and painting.

3.2 Clinical imaging study grouping

Slightly different debiasing trends for deep learning models have been observed for medical imaging – two key themes emerged from our review - (1) *multi-site* and (2) *demographic* fairness.

3.2.1 Multi-site fairness

Multi-site fairness focuses on model generalizability across different institutions/sites and remains a highly challenging issue for any model developer, especially since multi-site data accessibility is still limited due to privacy policies. Lack of generalizability at the multi-institutional level is multi-faceted, involving various shifts in data such as demographics, imaging systems, and imaging acquisition and reconstruction protocols. [36] constructed a model for detecting age on brain MR images and identified that models could trivially assess the site of origin of the data with an accuracy of 96% and embeddings generated by the model also demonstrated clear separation based on scanner subtypes. In literature, mostly optimal combinations

of training datasets were employed for increasing variations in the training data to reduce institutional bias and achieve multi-site fairness. However, [36] used two additional losses to unlearn the domain features from the MR images - (1) *domain loss* to assess how much domain information remains; (2) *confusion loss* that removes domain information by penalizing deviations from a uniform distribution, and were able to improve classification performance by domain adaptation where they removed confounding factors by creating a feature space that is invariant to the acquisition scanner.

[37] utilized two chest X-ray datasets (with tuberculosis) from the Montgomery County, USA and Shenzhen, China and showed that the model failed to efficiently generalize on unseen populations. A separate model with same architecture, trained on a mixture of the two datasets, achieved the highest overall performance. The authors concluded that robust AI models can be built using larger cross-population datasets. Similarly, [38] studied models ability to detect pneumonia in chest radiographs by training and testing on data from different hospital systems in the US - National Institutes of Health Clinical Center (NIH; 112,120 from 30,805 patients), Mount Sinai Hospital (MSH; 42,396 from 12,904 patients), and Indiana University Network for Patient Care (IU; 3,807 from 3,683 patients). The prevalence of pneumonia was high enough at MSH (34.2%) relative to NIH and IU (1.2% and 1.0%) that merely sorting by hospital system achieved an AUC of 0.861 (95% CI 0.855–0.866) on the joint MSH–NIH dataset. Utilizing a training set with equal incidence across sites achieved the best performance on the testing set, suggesting balanced disease prevalence allows the model to learn disease specific features instead of site-related features.

[39] dealt with three types of common data related biases for histopathology slides - (1) generates an allegedly good training performance whereas performances on independent datasets will be poor, (2) correlated with class labels and (3) sampling biases. They developed a model interpretation with pixel-wise heatmaps based on Layer-wise Relevance Propagation (LRP) [40] technique and visually evaluated potential dataset biases responsible for poor generalization for elimination of the biased data points from the training data incrementally.

3.2.2 Demographic fairness

One key unresolved question related to patient Health Insurance Portability and Accountability Act (HIPPA) is whether it is possible to re-identify patients based on their clinical information. Re-identification from data is not just a theoretical concept but has been demonstrated in several con-

texts, including clinical and social history datasets. For instance, demographics in an anonymized data set can function as a quasi-identifier that is capable of being used to re-identify individuals [41]. The risk of patient re-identification and resultant privacy violation related legal ramifications can further increase the reluctance of allowing AI models access to comprehensive clinical datasets further hindering fair AI model development and evaluation. Demographic fairness focuses on studying model performance for a single patient population with diverse racial/ethnic subgroups. By understanding demographic shifts, it is feasible to de-bias models at both - local and multi-institutional levels.

The greatest burden of biased models are felt by minority subgroups for whom the model may fail to diagnose accurately. [42] studied the impact of demographic biases in chest x-ray disease classification models using three large publicly available datasets. They quantified performance disparities by observing differences in true positive rates (TPR) among different protected demographic attributes such as patient sex, age, race, and insurance type which could be individually or in combination highly correlated with socioeconomic status. For example, women were found to have significant differences in TPR despite contributing to 50% of the dataset. Further analysis of disease prevalence found varying levels of disparities in disease where women had an equal proportion of disease positive cases.[43] investigated the impact of demographics on model performance while studying the classification of retinal fundus images for identification of age-related macular degeneration. The group used images from 68,400 individuals with available fundus images in the UK Biobank with additional manual classification in a subset of 2,013 participants, and found consistent misclassifications related to physical characteristics where light eyes were found to have a 1.2% false positive rate whereas darker eye color and darker fundus images had a larger false positive rate of 3.0% when used for classification tasks.

Following the proposed methodologies for natural images re-ID task, [44] studied the trade-offs between privacy-preserving techniques, and demonstrate that differential privacy techniques (DP-SGD [45] - the de-facto approach for linear models and neural networks) might not carry over to large medical datasets. When applied to the NIH Chest X-Ray dataset, many positive disease labels were eliminated in the training process to preserve privacy, and the removal of samples severely impacted model performance, reducing AUC from an average of 0.84 to 0.51. They also demonstrated that DP-SGD loses important information about minority classes (e.g., dying patients, minority ethnicities) that lie in the tails of the distribution. [46] studied the impact of gender balance in medical imaging datasets used to train AI

systems, and experimented with three deep neural network architectures and two well-known publicly available X-ray image datasets. The study shows that gender imbalance in chest x-ray datasets produces biased classifiers for diagnosis based on all models, with significantly lower performance in underrepresented groups. Interestingly, they did not find significant differences in performance between models trained with a gender-balanced dataset and an extremely imbalanced dataset from same gender which suggests diversity of demographic attributes is important for improved overall model performance.

4 Discussion

In this work, we performed a systematic review on 'Fair' AI models proposed for natural and medical images. Despite the large body of work focusing on AI model disparities in justice, law and education, there is little analytic work describing disparities in imaging applications. We screened 633 manuscripts, however only 22 relevant manuscripts were included in our analysis as mostly studies analyze the bias in the AI models without proposing any methodological solution. Among the 22 manuscripts, 14 manuscripts proposed solutions for natural images and focused two core themes - 'fair' training dataset (4/14) and representation learning (10/14). In debasing of medical image analysis, our review yielded 8 relevant manuscripts that carried two prevailing themes for medical images: model disparity across institutions (5/8) and model fairness with respect to patient demographics (3/8).

Debiasing techniques for natural images primarily focused on learning 'less bias' representation of the images for either classification of facial characteristics, age estimation, or developing a camera agnostic model for various image recognition tasks. Researchers have primarily explored three methodologies - (i) adversarial learning, (ii) joint learning and unlearning algorithm inspired in domain and task adaptation methods, and (iii) regularization of loss based on mutual information between feature embeddings and bias. Privacy-preserving approaches have also been proposed to disentangle certain attributes from learned representations. Clever sampling techniques and target task adjustment have also been proposed for reducing implicit bias in the training data.

While methods for natural images primarily focused on improving the three broad fairness criteria - demographic parity, equality of odds, and equality of opportunity - the primary aim of medical image debiasing is

to improve cross-institutional fairness which involves understanding model generalization under large shifts in patient and imaging distributions. Not all institutions have the resources to train a state-of-the-art deep neural network for diseases prediction, and current business models treat algorithms as medical products that can be sold across health systems. Studies evaluating models on external datasets demonstrate consistent drops in performance. Inspection of these failures reveal models learn feature associations that are disease prevalence-dependent and site- specific.

Even within a health system, there are several sources of variability (e.g. scanner type and patient demographics). From a clinician's perspective, detection of disease should primarily focus on the pathology present in an image. Yet, several papers have demonstrated that models can learn spurious associations between features in classification and prediction tasks. Other works have attempted to hide demographic information from model predictions with dataset and/or model optimization, often producing inferior performance. Simpler approaches such as balancing training set across demographics improve cross-group performance. However, model performance on tests with equal proportions found bias across classes.

In an attempt to explore a relatively simpler way to improve model performance, several studies suggested that having a greater number of positive cases across demographics helped models perform better in validation. When evaluating the resulting performance, AUC is a good overall measure but does not accurately capture precision and recall within subpopulations. Newer metrics such as True Positive Rate (TPR) gap could provide a better indication of a model's impact to patient care as it could be a surrogate for inherent model bias. Popular approaches to remove biases such as building demographic-specific models often suffer from a lack of demographic representation. The main challenge and goal of model debiasing is to reduce bias without the need for demographically balanced datasets. We observed that techniques that attempt to decouple demographic information and task predictions are not able to match baseline model performances and algorithmic development in medical imaging field is limited (Fig. 2).

Because fairness in healthcare is a social problem, there is no simple solution or "silver bullet" that can "solve" fairness in AI systems. However, AI can play a role in developing fairer systems by reducing clinical decisions that are tainted by unconscious or conscious bias. Model development must reflect learned disease features and not patient demographic or imaging protocol to ensure robustness across diverse populations. Further exploration of demographic/institution agnostic models would be worthwhile to minimize disparity gaps. Once models are built, the second challenge is the report-

ing of performance metrics that estimate its impact on outcomes among marginalized groups in addition to accuracy. The use of the TPR gap and FPR comparisons may provide better insight into how the model may perform on different patient subgroups. AUC provides a good metric for overall model performance, but provides little intuition on how AI may widen the gap in outcomes across populations. More importantly, the current evaluation of fairness in AI models is still mainly focused on accuracy based on real-world data. If all we are aiming for is accuracy of predictions and classifications across populations, then the future world will be no different from today - filled with inequities. If we want to use AI to prevent perpetuating or even magnifying health disparities, we need to come up with metrics that are not solely based on accuracy. How do we train and evaluate models when the ground truth is not fair? These are questions that will require the machine learning community to work more closely with social scientists.
Limitations: Our systematic review is limited due to restriction to the imaging domain (and not general healthcare AI) and given our focus on papers that describe novel methodology for debiasing, we were limited to only examining a small set of eligible research studies. There exists a larger field of bias and fairness work that was excluded due to not being related to deep learning or fitting to our criteria. The topic of fairness itself is relatively new to AI applications, limiting our review to works published in the last five years. More research works may exist in the field through pre-prints that were excluded due to lack of peer review. Further analysis of biases was limited by papers only presenting overall performance metrics instead of group specific performance.

5 Data availability statement

The authors declare that all data supporting the findings of this study are available within the paper. Additional review data can be shared upon request in Covidence.

References

[1] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning,"

[2] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. A. Finkelstein, E. L. Lamoureux, I. Y. Wong, N. M. Bressler, S. Sivaprasad, R. Varma, J. B. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," vol. 318, no. 22, pp. 2211–2223.

[3] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss, "Deep learning in mammography: Diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer," vol. 52, no. 7.

[4] H. Lee, E.-J. Lee, S. Ham, H.-B. Lee, J. S. Lee, S. U. Kwon, J. S. Kim, N. Kim, and D.-W. Kang, "Machine learning approach to identify stroke within 4.5 hours," *Stroke*, vol. 51, no. 3, pp. 860–866, 2020.

[5] L. Seyyed-Kalantari, H. Zhang, M. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature medicine*, vol. 27, no. 12, pp. 2176–2182, 2021.

[6] R. B. Parikh, S. Teeple, and A. S. Navathe, "Addressing bias in artificial intelligence in health care," vol. 322, no. 24, p. 2377.

[7] M. Whittaker, M. Alper, O. College, L. Kaziunas, and M. R. Morris, "Disability, bias, and AI," p. 32.

[8] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[9] R. Benjamin, "Assessing risk, automating racism," *Science*, vol. 366, no. 6464, pp. 421–422, 2019.

[10] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, and M. Ghassemi, "Hurtful words: Quantifying biases in clinical contextual word embeddings," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, (New York, NY, USA), p. 110–120, Association for Computing Machinery, 2020.

[11] A. S. Adamson and A. Smith, "Machine Learning and Health Care Disparities in Dermatology," *JAMA Dermatology*, vol. 154, pp. 1247–1248, 11 2018.

[12] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classication," p. 15.

[13] I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, P.-C. Kuo, M. P. Lungren, L. Palmer, B. J. Price, S. Purkayastha, A. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R. Wang, Z. Zaiman, H. Zhang, and J. W. Gichoya, "Reading race: Ai recognises patient's racial identity in medical images," 2021.

[14] C. J. Wallis, A. Jerath, N. Coburn, Z. Klaassen, A. N. Luckenbaugh, D. E. Magee, A. E. Hird, K. Armstrong, B. Ravi, N. F. Esnaola, *et al.*, "Association of surgeon-patient sex concordance with postoperative outcomes," *JAMA surgery*, vol. 157, no. 2, pp. 146–156, 2022.

[15] A. Kaushal, R. Altman, and C. Langlotz, "Geographic distribution of us cohorts used to train deep learning algorithms," *Jama*, vol. 324, no. 12, pp. 1212–1213, 2020.

[16] S. E. Davis, R. A. Greevy Jr, T. A. Lasko, C. G. Walsh, and M. E. Matheny, "Detection of calibration drift in clinical prediction models to inform model updating," *Journal of biomedical informatics*, vol. 112, p. 103611, 2020.

[17] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *BMJ*, vol. 339, 2009.

[18] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, pp. 1521–1528, IEEE, 2011.

[19] Z. Kou, Y. Zhang, L. Shang, and D. Wang, "Faircrowd: Fair human face dataset sampling via batch-level crowdsourcing bias inference," in *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pp. 1–10, IEEE, 2021.

[20] A. Clapes, O. Bilici, D. Temirova, E. Avots, G. Anbarjafari, and S. Escalera, "From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation," in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2373–2382, 2018.

[21] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pp. 1–7, IEEE, 2017.

[22] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[23] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "Sensitivenets: Learning agnostic representations with application to face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2158–2164, 2020.

[24] H. Zhang, H. Cao, X. Yang, C. Deng, and D. Tao, "Self-training with progressive representation enhancement for unsupervised cross-domain person re-identification," *IEEE Transactions on Image Processing*, 2021.

[25] Z. Alsulaimawi, "Variational bound of mutual information for fairness in classification," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, 2020.

[26] N. Quadrianto, V. Sharmanska, and O. Thomas, "Discovering fair representations in the data domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8227–8236, 2019.

[27] L. Jiang, J. Zhang, and B. Deng, "Robust rgb-d face recognition using attribute-aware loss," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2552–2566, 2019.

[28] E. Adeli, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, J. C. Niebles, and K. M. Pohl, "Representation learning with statistical independence to mitigate bias," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2513–2523, 2021.

[29] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European conference on computer vision*, pp. 17–35, Springer, 2016.

[30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.

[31] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 4, pp. 956–973, 2018.

[32] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, vol. 3, pp. 1–7, Citeseer, 2007.

[33] L. Yan, R. Zhu, N. Mo, and Y. Liu, "Cross-domain distance metric learning framework with limited target samples for scene classification of aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3840–3857, 2019.

[34] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised domain adaptation for depth prediction from images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2396–2409, 2019.

[35] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

[36] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal," vol. 228, p. 117689.

[37] D. Das, K. C. Santosh, and U. Pal, "Cross-population train/test deep learning model: Abnormality screening in chest x-rays," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 514–519. Journal Abbreviation: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS).

[38] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study.," vol. 15, no. 11, p. e1002683.

17

[39] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," vol. 10, no. 1, p. 6423.

[40] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[41] L. Sweeney, "Achieving ¡i¿k¡/i¿-anonymity privacy protection using generalization and suppression," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, p. 571–588, Oct. 2002.

[42] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest x-ray classifiers,"

[43] F. Guenther, C. Brandl, T. W. Winkler, V. Wanner, K. Stark, H. Kuechenhoff, and I. M. Heid, "Chances and challenges of machine learning-based disease classification in genetic association studies illustrated on age-related macular degeneration.," vol. 44, no. 7, pp. 759–777. Place: United States.

[44] V. M. Suriyakumar, N. Papernot, A. Goldenberg, and M. Ghassemi, "Chasing your long tails: Differentially private prediction in health care settings," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 723–734, Association for Computing Machinery. event-place: Virtual Event, Canada.

[45] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[46] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 23, pp. 12592–12594, 2020.

[47] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," vol. 117, no. 23,

pp. 12592–12594. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/117/23/12592.full.pdf.

6 Authors Contribution

Concept and design: R.C., J.G., and I.B. Study selection: R.C., M.S and I.B. Data extraction: R.C., M.S and I.B. Drafting of the manuscript: R.C., M.S, J.G., B.P., and I.B. Critical revision of the manuscript for important intellectual content: H.T, B.P, L.C., Supervision: I.B.

7 Competing Interests statement

Authors declare no conflict of interest.

19

8 Figure

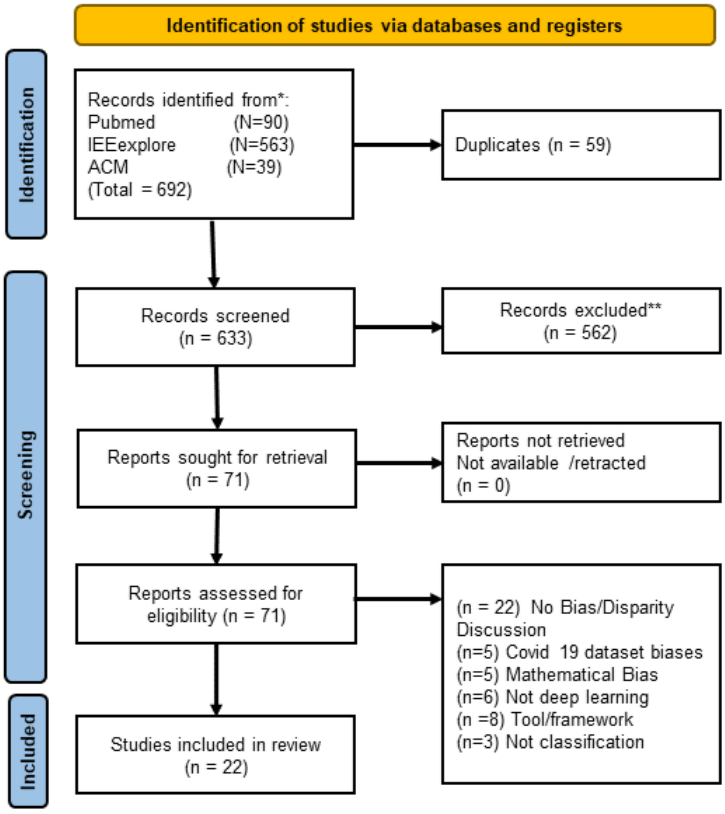

Figure 1: PRISM flowchart for literature screening. n represents number of unique papers.

20

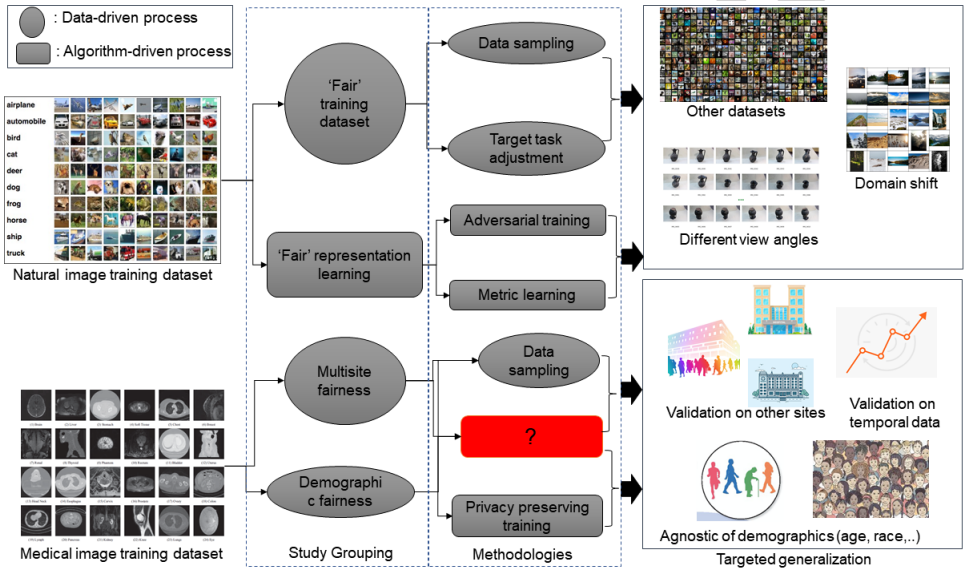

Figure 2: Pictorial representation of study grouping strategy: natural images and medical images.

21

9 Tables

22

Authors & publication date	Dataset origin	Modality (type)	Method for De-biasing	Targeted task	Performance	Observations
[18]	KITTI Raw Dataset	LIDAR images	Fine-tuning deep learning model using errors weighted by confidence score of smaller model	Depth Prediction	No Adaptation - target domain bad3: 10.86 - target domain mae: 1.73 -similar domain bad3: 0.86 -similar domain mae: 1.73 Complete Adaptation - target domain bad3: 2.96 - target domain mae: 0.96 -similar domain bad3: 3.66 -similar domain mae: 1.04	Smaller models consistent errors can be exploited to tune the learning process
[19]	CelebA dataset	Face Images	Sampling fair sub-dataset by leveraging inferred demographic labels through a novel shuffling process	Face Attractiveness Prediction	0.102% lower demographic Parity 0.292% lower Equalized Odds	Efficient sampling can improve model fairness and performance.
[20]	APPA-REAL dataset	Person Images	Utilize subgroup's average apparent age as a correction of true age estimation	Age estimation	Baseline MAE: 13.56 Correction on Gender: 11.91	Demographic biases can be exploited to reduce error
[21]	Child emotions dataset: NIMH Dartmouth Radbound CEP	Person Images	Downstream model for prediction between two ambiguous classes	Emotion prediction	Baseline accuracy: Dartmouth: 25% Radbound: 33% CEP: 5% After debasing: Dartmouth: 70% Radbound: 71% CEP: 55%	Additional downstream models are useful for improving class separation

Table 1: Benchmarking articles related to proposing new dataset sub-sampling techniques for reducing bias.

23

Authors & publication date	Dataset origin	Modality (type)	Method for De-biasing	Targeted task	Performance	Observations
[23]	Dive Face dataset	Facial Images	Adversarial regularizer.	Face identification	Adverse task: Gender: before 97.7, after: 58.8 Ethnicity: before 98.8%, after 55.1%	Privacy preserving representations learning for bias reduction.
[25]	CelebA	Facial images	Feature protection using adversarial training	Smile detection	Target: AUC 0.97 Advers$^{(}*$): AUC: .51	Autoencoders can remove irrelevant features allow training a fair classifier.
[26]	CelebA, DiF	Facial images	Unconstrained mapping using residual decomposition	Face attractiveness prediction & Age prediction	CelebA Baseline acc:80.6 Proposed: acc:79.4	Removal of biased features can reduce biased outcomes.
[24]	Market1501, DukeMTMC-reID, MSMT17	Person Images	Unsupervised multi-scale self-training for view-invariant representation learning	Person Re-identification	Dataset increase: avg 3.16% mAP, 5.16% in rank accuracy	confounding features can be separable during adversarial debiasing.
[31]	VIPeR ,CUHK0, CUHK0, SYSU, Market-1501, ExMarket	Person images	Asymmetric metric learning	Person identification	ExMarket: -DIC model MAP 42.18 (21.19) -DECAMEL model 62.98 (33.28)	Effect of task irrelevant data clusters can be reduced using distance-based loss.
[27]	Private Dataset	Person Images	Attribute aware loss	Face identification	Accuracy RGB: 87.50 RGB+Depth: 92.84 RGB+Depth+attribute: 96.5	Task performance can improve by including demographic attributes in training.
[33]	WHU-RS,SIRI-WHU,jinmen	Arial images	image to subcategory distance function	classification of land cover	Baseline (res): 92.99, 88.73, 87.01 Proposed : 96.69, 94.50, 91.74	Additional loss can enforce similarity across domains.
[28]	Private Brain MRI dataset Gender Shades Pilot Parliaments Benchmark	Brain MRI Person Images	Domain adaptation via statistical independence	Prediction of HIV status Gender prediction	MRI (AUC,corr) : (80.9,.05) Gender: BR-NET: 99.4,.12	Learn better representations via reduced correlation between features & confounding task
[34]	KITTI Raw Dataset	LIDAR images	Domain adaptation via model fine-tuning	Depth Prediction	Baseline MAE 1.73 Complete Adaptation MAE:0.96	Related task can provide additional error estimation during the adaptation process
[35]	PACS (Photo, Art , Cartoon, Sketch), VLCS dataset	Photo, Art, Cartoon, Sketch	Domain generalization based on low-rank parameterized CNNs	Non photo-realistic image classification	VLCS proposed MAP: 72.11% Baseline MAP 72.02% PACS proposed MAP: 69.21% Baseline: 67.37%	Learning a subset of parameters to be domain specific improves generalization

Table 2: Benchmarking articles related to proposing new techniques for 'fair' representation learning from generic images.

Authors & publication date	Dataset origin	Modality	Method for de-biasing	Targeted task	Performance	Observations
Institutional Bias						
[39]	TCGA SKCM TCGA-BRCA TCGA-LUAD -open source	Digital Pathology	Remove biasing data samples	Cancer cell detection	Improve AUC by 0.05 with de-biasing	Model interpretation technique to discover hidden bias.
[37]	Montgomery County Hospital, USA & Shenzhen Hospital China - private	Chest X-ray	Dataset consolidation	Predicting chest X-ray findings	AUC:0.84	Performance improved across sites by balancing prevalence of diseases.
[38]	Mount Sinai Hospital and Indiana University, USA - Private	Chext X-ray	Dataset balancing with equal site incidence	Diagnosis of abnormalities in chest X-ray	AUC:0.815	Balancing proportions of site incidence achieved improved performance
[36]	UK BioBank OASIS dataset (healthy only) Whitehall II Study (healthy Only) - open-source	Brain MRI	Confusion Loss	Age Estimation	Dataset MAE:(Original,Unlearning) Biobank: 3.24, 3.38 OASIS: 4.19,3.90 Whitehall: 2.89 ,2.56 Scanner Accuracy: 96,34 %	Separation of correlated task and confounder features through selection of data subsets.
Demographic Bias						
[44]	NIH-Chest X-ray - open-source	Chest X-ray	Privacy preserving training	Predicting chest x-ray findings	AUC:0.84	Differential privacy techniques suffer performance decrease despite theoretical guarantees.
[47]	NIH Chest X-ray - open-source	Chest X-ray	Balancing proportions for demographic factors	Diagnosis of abnormalities in chest X-ray	AUC: 0.81	Training on equal proporotion of gender improves model performance.
[42]	NIH Chest X-ray, MIMIC chexpert - open-source	Chext x-ray	Observational, no debiasing	Diagnosis of abnormalities in chest X-ray	Sex TPR GAP: .045 Age TPR GAPR:.215 Race TPR GAP: .226	Positive correlation exist between diseases incidence and demographic disparity but not statistically significant
[43]	EU BIOBank - open-source	Retinal fundus	Observational, no debiasing	Age related macular degeneration classification	%FP light eyes: 1.2 , %FP in dark eyes: 3.	Statistical analysis of misclassification; identified darker eye image to be misclassified more often.

Table 3: Benchmarking articles related to proposing new debiasing techniques for medical imaging.

25