A Survey of Fairness in Medical Image Analysis: Concepts, Algorithms, Evaluations, and Challenges

Zikang Xu, Jun Li, Qingsong Yao, Han Li, Xin Shi and S. Kevin Zhou*, Fellow, IEEE

Abstract—Fairness, a criterion focuses on evaluating algorithm performance on different demographic groups, has gained attention in natural language processing, recommendation system and facial recognition. Since there are plenty of demographic attributes in medical image samples, it is important to understand the concepts of fairness, be acquainted with unfairness mitigation techniques, evaluate fairness degree of an algorithm and recognize challenges in fairness issues in medical image analysis (MedIA). In this paper, we first give a comprehensive and precise definition of fairness, following by introducing currently used techniques in fairness issues in facial recognition. After that, we list public medical image datasets that contains demographic attributes for facilitating the research on fairness in MedIA, and summarize current research on fairness in MedIA. TO help achieve a better understanding of fairness, and call attention on fairness related issues in MedIA, experiments are conducted comparing the difference between fairness and data imbalance, verifying the existence of unfairness in MedIA, especially in classification, segmentation and detection, and evaluating the effectiveness of unfairness mitigation algorithms used in facial recognition on MedIA. Finally, we conclude with opportunities and challenges in fairness in medical image analysis.

Index Terms—Fairness, Medical imaging, Survey

I. INTRODUCTION

A. Background

The past years have witnessed a surge in deep learning (DL) methods, for their prevalence performance in computer vision (CV) and nature language processing (NLP) tasks. DL has become an integral part in systems like automatic driving, facial recognition and healthcare analysis. Specifically, DL is maturing at a fast pace in medical research and applications, from low-level tasks, like image reconstruction, denoising and enhancement, to high level tasks including classification, segmentation and detection [1].

While current research is mostly concerned about deriving higher performance throughout the evolution of DL algorithms, such as the prediction accuracy on classification tasks, the Dice similarity score on segmentation tasks, etc., there is a growing interest in going beyond mere performance metrics by addressing, say, the interpretability, explainability and trustworthiness aspects of DL methods. The eXplainable Artificial Intelligence (XAI) methods concern on explaining their rationale, characterizing their relative merits, and conveying an understanding of how they will behave in the future. In general, XAI focuses on explaining algorithm, including trustworthy, interpretability, fairness, causality, reproducibility, etc. [2]. In the FUTURE-AI initiative toward trustworthy medical image analysis [3], six guiding principles are presented: Fairness, Universality, Traceability, Usability, Robustness and Explainability. Fairness, as the first and the fundamental principle, calls upon the importance of considering fairness when designing algorithms in medical applications.

Concerns on the fairness of DL model derived from a prominent debate between Yann Lecun and Timnit Gebru in 2020. It all started from a super resolution algorithm PULSE [4], which was accused for ‘racism’. PULSE is designed to improve the image resolution, and generate a rich detailed and high fidelity image. Surprisingly, people found that a portrait image of Barack Hussein Obama got a clearer output image with a white man’s face, after processing by PULSE. Although the reason may be the imbalance distribution of training data, this still caused an extensive concern on the bias and fairness issues in DL models.

As is widely recognized, fairness presupposes that all people have the same right and should be treated equally, which means DL models should not be discriminatory on factors like race, colour, sex, language, religion, political, national, property or birth. The issue of fairness brings longstanding discussions on AI transparency and algorithmic non-discrimination in company promotion [5], university entrance evaluation [6], music recommendation [7], etc.

There is a consensus in the research of fairness that the meaning of fairness is not analytically well-defined, and several definitions are even orthogonal [8]. For example, different fairness criteria require different constrains on the model’s performance [9]. Among all these fairness definitions, two commonly used definition are individual fairness [10] and group fairness [11]. Individual fairness requires that similar individuals should be treated equally and thus have similar predictions. For example, a model should have comparable diagnosis on two similar X-Ray images (the concept ‘similar’ is defined using some similarity measurements including cosine-similarity [12]. While group fairness requires equal performance on the groups divided based on sensitive attributes (e.g., race, sex, and age) first. In image processing area, group
fairness is more frequently used than individual fairness, and in this survey we also focus on group fairness.

Recently, researchers pay attention to fairness issues in medical applications. Comparing with other areas, fairness in medical image analysis is more valuable and important, because the relationship between algorithm and human is closer due to the widespread use of medical images in clinical decision making.

It is clear that the evaluation of fairness of deep learning algorithms is, and will continue to be, a common scenario and inevitable issue in training deep learning models. Hence, the concepts of fairness, the relationship between fair medical image analysis and fairness issues in other fields, current research in medical image analysis and potential analysis and research directions and challenges are highly desired. However, there is a lack of systematic review of fairness in medical image analysis. In this paper, we attempt to bridge this gap.

B. Aims and scope

We first provide a precise and detailed definition of group fairness and compare it with similar concepts. We then review the methods proposed in facial recognition, most of which have not been widely applied in medical image analysis. After that, we introduce the collection of medical image datasets that can be used in fairness analysis in medical applications, including the modality and sensitive attributes. Then, we present our experimental on several datasets, where we investigate the importance and necessity of fairness evaluation and mitigation in medical applications, and assess the effectiveness of fairness mitigation methods. Finally, we propose the ongoing difficulties and challenges in fair medical analysis and point out several potential research directions in this area.

The structure of this paper is as follows: Chapter II introduces basic concepts of fairness and compares fairness with other related concepts; Chapter III presents current research of fairness in facial recognition; Chapter IV surveys current research in fairness in medical image analysis; Chapter V evaluates the value and necessity of fairness evaluation in area of medical image analysis by several experiments; and Chapter VI offers the opportunities and challenges in fair medical image analysis and concludes the whole survey. Fig. 1 provides a summary of paper organization.

II. CONCEPTS OF FAIRNESS

In order to analyze fairness, we need to have a clear and precise definition of fairness. “Fairness” remains a complex and controversial concept in machine learning and deep learning area until Narayanan et al. [13] gives a systematic and formulaic definition.

Unlike other evaluation metrics of DL algorithms, fairness criterion is a metric that focuses the relationship between algorithm performance and human factors. Take a simple classification task as an example, the information of patient can be separated as task-related information (medical images, denoted as $X$), e.g., MRI images and X-ray images, and task-irrelevant information that is inherent, which is called sensitive attributes, $A$, such as age, gender, race, etc. These sensitive attributes can separate patients into several groups, like male / female, Asian / African / American. We need to emphasize here that, although in some scenarios, sensitive attributes are related to target task, we still regard these attributes as task-irrelevant and do not expect to categorize the classification task via this information, because due to the distribution of dataset, algorithms prefer to use the easiest distribution of dataset, algorithms prefer to use the easiest shortcut learning (for example, classify all female patients as with illness A and all male patients as with illness B). Besides, we temporarily suppose that the medical images do not include any information of sensitive attributes (in fact, sensitive attributes like gender can be extracted from medical images like [14]. Therefore, we can use a casual graph to describe this classification task as in Fig. 3a.

In order to achieve algorithmic fairness, we want to break the connection between sensitive attributes $A$ and target task output $T$ (see right subfigure in Fig. 3a). In other words, absolute fairness means that algorithm has the same performance on different groups of patients, regardless their different sensitive attributes.

A. Definition of fairness

Suppose we have a dataset $D$ with $N$ samples $d_1, d_2, \ldots, d_N$, the $i$-th sample $d_i$ consists of medical image data $X_i$, sensitive attributes $A_i$ and target task ground truth label $Y_i$, i.e. $d_i = \{X_i, A_i, Y_i\}$. A medical image $X$ is an arbitrary-dimensional array depends on specific task, sensitive attributes $A = \{A^1, A^2, \ldots, A^L\}$ contain $L$ elements, with each sensitive attribute $A^l$ representing a discrete variable (e.g., gender) or a continuous variable (e.g., age). For processing convenience, we discretize a continuous sensitive attribute. For a discrete variable $A^l$, we denote the number of possible values by $C_l$. Therefore, the whole dataset $d$ is split into $G = \prod_{l=1}^L C_l$ groups, each group has $N_g$ samples with $\sum_{g=1}^G N_g = N$.

For a typical machine learning model (e.g., a neural network) $f$, it takes $X$ as the input and outputs the prediction of
$Y$, $\hat{Y}$. We use a distance criterion $M$ to evaluate the difference between $Y$ and $\hat{Y}$, that is,

$$\hat{Y}_i = f(X_i),$$

(1)

$$D_i = M(Y_i, \hat{Y}_i).$$

(2)

Then, for a group $g$, $g \in [1, G]$, we compute its performance $D^g$ by simply averaging the distances of all samples in group $g$, that is,

$$D^g = \frac{1}{N_g} \sum_{i=1}^{N_g} D_i,$$

(3)

Therefore, absolute fairness means that $D^1 = D^2 = D^3 = \cdots = D^G$. Once the criterion $D^g$ is not the same, first-order or higher-order statistics is computed for unfairness evaluation. In this paper, we use privileged to refer to groups that have performance higher than average, and unprivileged to refer to groups that have performance lower than average. Details about how to measure the degree of unfairness can be found in Section II-E3.

B. Type of sensitive attributes

From the about description, it is clear that, sensitive attributes include any information that can be extracted from patients. For categorical attributes, we can use it to separate the whole dataset easily. However, if sensitive attribute is a continuous variable, we need extra processing before fairness analysis, for example, we can factitiously divide age into young group and elder group using a threshold age considering the dataset distribution.

From the aspect of the relationship between sensitive attributes and patient, we can categorize sensitive attributes into congenital and postnatal. More attention should be paid when dealing with postnatal attributes, since the definition of postnatal attributes may introduce noise for fairness evaluation. For instance, in some cases, patients that drink more than twice a week are regarded as bibulosity while in other cases, only patients who drink more than 5 times weekly are regard as bibulosity.

In Fig. 2, we list the sensitive attributes commonly used in medical applications and categorize them from the aspect of value type and the aspect of relationship to patient.

C. Similarity and difference between similar concepts

For a clear understanding of fairness issue and distinguishing fairness from similar concepts, we below compare fairness with data imbalance, domain adaptation, treatment effect estimation and privacy preserving, which all have a close relationship with fairness.

Data Imbalance is a common issue in deep learning research, which refers to imbalanced distribution of values of the response variable [15]. There are two major differences between fairness and data imbalance. First, data imbalance concentrates on the difference in the distribution of ground truth labels of target task, while fairness focus on the difference in the performance of the groups with different sensitive attributes. Second, it is not clear whether data imbalance is the source of unfairness. According to the experiment results in V-A, we can find that even the dataset is balanced, unfairness still exists.

Domain Adaptation is another concept that is similar with fairness. Usually, there are two (or more) domains in a typical domain adaptation, source domain $S$ and target domain $T$ [16], which is akin to group of different sensitive attributes in fairness. In domain adaptation, we aim to improve the performance in target domain, making it approximate to the performance in the source domain. Similarly, in fairness, we want the groups with different sensitive attributes to have the same performance. However, there are two major differences between domain adaptation and fairness. First, there is a clear primary and secondary relationship between source domain and target domain, while in fairness, there is no explicit privilege relationship between groups with different gender; Second, in fairness issue, usually we can mitigate unfairness by degrading the performance of privileged group. However, in domain adaptation, the performance of source domain is fixed, which means that we can regard fairness as a weaker task of domain adaptation to some extent.

Treatment Effect Estimation, another similar concept to fairness, is a measure used to compare treatments in randomized experiments and medical trials by measuring the difference in average outcomes between units assigned to the treatment and units assigned to the control [17]. Fig. 3 shows the comparison of a typical causal graph of fairness and treatment effect estimation. From the figure we can find that in treatment effect estimation, the target is to establish the relationship between binarized treatment $T$ and output $Y$, and the confounder is electronic health records (medical image data). While in fairness, the target is to establish the relationship between medical image data $X$ and output $Y$, and the confounder is a sensitive attribute, which is reflected...
in the results, having the same performance on the groups with different sensitive attributes. Although the dimension of treatments and medical image data vary considerably, we can find that the causal structure and logical relationship of these two tasks are the same. Therefore, we hold the opinion that treatment effect estimation is a kind of fairness task.

Privacy Preserving is an important concern of recent interest. Dataset may include unwanted or sensitive information about the source of data. For example, sex information of facial images (CelebA Dataset [18]) and the spot for photography of street images (Cityspace Dataset [19]). Therefore, algorithms that try to hide private information are proposed. One of the typical privacy preserving method is Federal Learning (FL) [20], which trains model on server using different databases on separated clients. We hold the opinion that federal learning is one of the directions for unfairness mitigation since the goal of these two tasks is to hide protected information in some contexts. Besides, considering a specific scenario in federal learning that each client only contains samples with the same sensitive attributes (for example, samples in Client A are all captured from male patients while Client B only contains samples of female patients). Such a scenario yields a federal learning method that is almost totally fair.

D. Sources of unfairness

In the former part, we give fairness a clear definition by formulas and comparison with similar concept. Here, we introduce potential sources of unfairness. According to [21], we categorize the sources into inherent sources (unfairness from data) and postnatal sources (unfairness from algorithm).

1) Unfairness from data: The inherent reasons of unfairness major come from four aspects: social bias, noisy annotation, data imbalance and data collection.

Social Bias is the first reason for unfairness and the hardest to solve, which usually comes from historical problems. Taking employment problem for example, some companies prefer to employ male staff rather than female staff [22]. Besides, in COMPAS dataset [23], we can find that black inmates have higher recidivism than white defendants from the 2-year follow-up study. Due to complex historical and environmental reasons, it is difficult for researchers to separate social bias from target task for fair algorithm development. Thus, this factor is not considered in this paper because social bias is more like an ethic problem rather than an algorithmic problem.

Noisy Annotation of datasets is another reason of unfairness. In recent years, with the surge in demand for annotated data, crowd-sourcing becomes a common method for data labeling [24]. However, since the quality of labeling by crowd-sourcing is uncontrollable, additional noisy annotation is generated, which does harm to fairness, e.g., many facial expression datasets contain significant annotation biases between genders, especially on happy and angry expressions, which cannot be mitigated by traditional methods [24]. In medical applications, it is common that different doctors have different annotations on same patient, which also produce noisy labels especially on difficult cases. When measuring oxygen saturation of patients with COVID-19 using pulse oximetry, pulse oximetry overestimates arterial oxygen saturation among Asian, Black and Hispanic patients compared with White patients [25], which not only validates unfairness among different races, but also influences the judgement of COVID-19 therapies.

Imbalanced Data distribution also causes unfairness, due to the characteristics of “data-driven” of machine learning methods, which leads to over (or under) expression of features of data and further causes unfairness.

Data Collection Webb et al. [26] find that racial bias may influence the signal when using a neuroscience equipment for physiological data collection, because people of different races usually have different skin conductance response and thus add unwanted information in electroencephalography data. Besides, a recent research on COVID-19 finds that the words used in medical texts are starkly different when describing dark-skinned population versus light-skinned populations, which although might have little influence on disease diagnosis, but shows an overall disappointing unfairness [27].

2) Unfairness from algorithm: In addition to unfairness due to the use of data, there also exist some postnatal reasons for unfairness, including network architecture and loss function. These factors constitute the focus of this paper.

Network Architecture The network architecture of deep learning is a postnatal source of unfairness. According to [28], in training step, network tries to optimize its parameters using the easiest feature, which usually builds unwanted connection between outputs and confounders (sensitive attributes). In other words, errors in feature extraction can lead to unfairness in target task performance. Experiments in [29] show that, even state-of-the-art open source gender classifier can be affected by skin type changes, thereby leading to unfair predictions.

Loss Function Since the degree of fairness is quantized by
E. Evaluation of fairness

In this part, we introduce methods for fairness evaluation. With loss of generality, we present the case of a binary single sensitive attribute, i.e., $A \in \{0, 1\}$. A convenient example is that the sensitive attribute is male or female. First, we list traditional fairness metrics that have ethical meanings, which is used in classification tasks. Then, we broaden the scope of fairness metrics in a classification task and other common tasks in medical applications. Finally, we introduce scores that measure the trade-off between fairness and performance of tasks. Notice that, although these metrics are defined using binary sensitive attributes, they can be expanded to multi-value or continuous sensitive attributes easily.

1) Traditional fairness metrics: Barocas et al. [9] define a number of fairness criteria which are firstly used for fairness evaluation on tabular datasets. To be consistent with its description, when introducing traditional fairness metrics, we suppose that target task is a binary classification task, i.e., $Y \in \{0, 1\}$, where $Y = 1$ represents positive result and $Y = 0$ represents negative result, we use $\hat{Y}$ to represent the model output.

Demographic Parity (DP) The criterion of demographic parity states that the model outcome should not be affected by any sensitive attribute, which is given by the following formula:

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1),$$

i.e., the ratio of patients who are positive should be the same in male patients group and female patients group. Fig. 4(a) shows a simple example that the model reaches demographic parity among male and female patients.

Equal of Odds This criterion requires the independence of the outcome of model and sensitive attribute. In other words, disparities in groups with different values of sensitive attributes should be completely justified by the outcome of model [30]. This can be expressed as follows:

$$P(\hat{Y} = 1 \mid A = 1, Y = y) = P(\hat{Y} = 1 \mid A = 0, Y = y),$$

where $y \in \{0, 1\}$. Among all the patients who are positive, the ratio of male patients who is predicted as cancerous should equal with the ratio of female patients who is predicted as cancerous. Fig. 4(b) shows a simple example that the model reaches equal of odds among male and female patients.

Equal Opportunity This criterion is a relaxed criterion of Equal of Odds [31]. Equal opportunity only concerns on negative samples, which is defined by following formula:

$$P(\hat{Y} = 1 \mid A = 1, Y = 1) = P(\hat{Y} = 1 \mid A = 0, Y = 1),$$

Fig. 4(c) shows a simple example that the model reaches equal opportunity among male and female patients.

Accuracy Equality This criterion requires the classification system to have equal misclassification rates across sensitive groups [32]:

$$P(\hat{Y} \neq y \mid A = 1, Y = y) = P(\hat{Y} \neq y \mid A = 0, Y = y).$$

That is, the ratio of misclassified patients in male group should equals that in female group. Fig. 4(d) shows a simple example that the model reaches accuracy equality among male and female patients.

2) Generalized fairness metrics: The above four fairness metrics are only applicable for classification tasks and are derived from a confusion matrix. Specifically, demographic parity is based on positive prediction rates (Predictive Positive / Total Population); equality of odds based on true positive rates (TPR) and false positive rates (FPR); equal opportunity based on false negative rate (FNR); accuracy parity based on accuracy (ACC). In this paper, we generalize the fairness metrics for classification tasks with several other statistics, such as Negative Predicted Value (NPV) and F1 score, derived from a confusion matrix.

Furthermore, there is no proper fairness measurement metrics for other image processing tasks like segmentation, detection, etc. We choose Dice Similarity Coefficient and Hausdorff Distance 95% as fairness metrics for segmentation tasks, and IOU and Average Precision for detection tasks. Notice that for other special-interest image tasks including registration and landmark detection, and some detection tasks that only concern False Positive Per Image (FUPI) [33], researchers can use commonly used metrics that can be averaged on groups as fairness metrics. The list of these generalized metrics is shown in Table I.

3) Overall unfairness measurement: After computing these fairness criteria for each sensitive groups, there is a need for measuring the overall degree of unfairness. The commonly used measurements including subtraction [34], division [34], standard deviation [35] and NR (Normalized Range) [36]. Let $M^g, g \in [1, G]$ represent an arbitrary fairness criterion of the $g$-th sensitive group, these three overall measurements are given by the following equations:

$$O_{\text{Subtraction}} = \max\{\{M^g\}\} - \min\{\{M^g\}\};$$

$$O_{\text{Division}} = \frac{\min\{\{M^g\}\}}{\max\{\{M^g\}\}};$$

$$O_{\text{STD}} = \sqrt{\frac{\sum_{g=1}^{G} (M^g - \bar{M})^2}{n - 1}};$$

$$O_{\text{NR}} = \frac{\max\{\{M^g\}\} - \min\{\{M^g\}\}}{\frac{1}{G} \sum_{g=1}^{G} M^g}.$$
is proposed in [40], which draws
uses numeric computation to evalu-
generally follows one of two research
focuses on the quality of training

Fig. 4: Graphical illustration of traditional Fairness Criteria. (a) Demographic Parity: Male ($\frac{6}{12} = \frac{5}{10}$). (b) Equal of Odds: for $Y = 1$, Male ($\frac{6}{3} = \frac{3}{5}$), for $Y = 0$, Male ($\frac{2}{3} = \frac{2}{5}$). (c) Equal Opportunity: Male ($\frac{2}{3} = \frac{2}{5}$). (d) Accuracy Equality: Male ($\frac{6}{10} = \frac{4}{10}$).

4) Fairness-Accuracy trade-off score: Previous researches have shown that, in order to improve the degree of fairness, the average performance of algorithm will decrease [37]–[39]. Therefore, when evaluating the performance of an algorithm, we must consider both its performance on target task and fairness criterion, that is, the trade-off between performance and fairness should be taken into consideration. The commonly used methods for performance-fairness trade-off analysis and fairness mitigation methods evaluation can be separated into curve-based and value-based.

Curve-based Method is proposed in [40], which draws a figure containing several algorithms. The horizontal axis represents one kind of fairness measurement (for example, demographic parity $\Delta DP$ among sensitive groups), while the vertical axis shows average performance on the whole dataset (for example, classification accuracy of skin lesion). The closer the curve is to the top left corner, the smaller trade-off it has. This method can only give qualitative results on different models. An example of curve-based method is shown in Fig. 5, since the curve in blue is the closest to top left corner, model A has the best performance-fairness trade-off.

Value-based Method uses numeric computation to evaluate the trade-off between fairness and performance. Dhar et al. [41] introduce a metric called bias performance coefficient (BPC) defined by the following equation:

$$BPC = \frac{Bias - Bias_{deb}}{Bias} - \frac{Acc - Acc_{deb}}{Acc},$$  

(12)

where $Bias$ and $Bias_{deb}$ represent the origin model that does not consider fairness and the improved model that considers fairness, respectively. $Acc$ and $Acc_{deb}$ represent average accuracies of these two models, respectively. BPC is like ‘normalized fairness-accuracy trade-off’ to some extent.

F. Unfairness mitigation methods

After evaluating the degree of unfairness in specific task, the key point is using improved method to mitigate unfairness as much as possible. According to [42], unfairness mitigation methods can be categorized into pre-processing method, in-

Fig. 5: Curve-based method for trade-off evaluation. Horizontal axis: absolute difference in demographic parity among demographic groups. Vertical axis: classification accuracy. The closer the curve is to the top left corner, the better performance-fairness trade-off it has.

processing method and post-processing method according to targeting stage.

Pre-processing Method focuses on the quality of training dataset, some researchers collect novel fair datasets, some researchers use additional external datasets, some researchers use GANs or VAEs to generate extra images with different sensitive attributes. Besides, sampling strategies are also usually applied for training a more fair model.

In-processing Method generally follows one of two research directions. The first is adversarial network architecture, which is derived from domain adaptation area that regards sensitive attribute as domain-specific label and tries to only use domain-irrelevant features for target task. Another direction is by introducing extra fairness constraints in the training algorithm [42].

Post-processing Method does not need to change or re-train pretrained models. This type of methods usually modifies the
output of a previously trained network on different sample groups to achieve a specific fairness metric.

III. FAIRNESS IN FACIAL RECOGNITION

Analyzing fairness issues has been a long-standing problem in deep learning [43], [44], from the fundamental research in tabular datasets [6], [45], [46], to higher level research including fairness in natural language processing [47]–[52], fairness in recommendation systems [53], [54], fairness in image caption [55], [56] and vision question answering [57], [58] and fairness in facial recognition etc.

Studies have shown the importance and necessity of paying attention on the fairness of deep learning models besides target task performance [59].

In this section, we focus on fairness researches in facial recognition tasks as it takes facial images as input, which is more similar with medical image analysis tasks. Besides, most of the analysis of fairness in top conferences in computer vision field including IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV) are set up in the context of face recognition tasks.

As mentioned before, the methods of unfairness mitigation can be categorized into pre-processing, in-processing and post-processing. In this section we start with the evaluation of fairness in facial recognition areas and categories methods used for unfairness mitigation by targeting stage. Tab. II shows all the methods we reviewed.

A. Fairness evaluation

Several researches concentrate on evaluating the existence of unfairness in facial recognition tasks. [60] constructs a novel benchmark for studying unfairness in visual recognition models by building a skewed dataset from the original dataset and considers mean per-class per-domain accuracy and bias amplification of [61]. They analyzes many fairness mitigation techniques thoroughly and highlights the effectiveness and shortcomings of these methods. Researchers in [62] hold the opinion that due to the diversity of image in datasets, directly evaluate fairness of algorithms might be disturbed. They develop an experimental method for measuring algorithmic unfairness of face analysis algorithms by generating synthetic image grids that along specific attributes while leaving other attributes constant using parameter controllable GAN. After re-annotating the synthetic images by human, they measure attribute-specific unfairness by comparing the algorithm’s predictions with human annotations.

Another routine of fairness evaluation focus on the image caption tasks that associated with human attributes. [56] studies unfairness propagation pathways within image caption on the COCO dataset, and find that one source of unfairness in image caption might comes from the crowdsourcing demographic annotations. They find that the performance of image caption models differs slightly between lighter and darker images, and even the vocabulary they use for describing people differs based on skin tone.

B. Unfairness mitigation

We split unfairness mitigation algorithms into three categories based on their applying stage, i.e., pre-processing, in-processing, and post-processing.

1) Pre-processing method: Fairness mitigation via pre-processing in facial tasks mainly consists of two folds. First, attribute-controllable image synthesis: [63] uses GANs to generate realistic-looking images and augments the original dataset by these images. By vector operations in the latent space, the synthesis images have the same target label with the origin image, while have the opposite sensitive attributes to the origin image. The experiments on augmented dataset shows both quantitative and qualitative improvements comparing with training with origin dataset. [64] also proposes a generalized variant of SMOTE (g-SMOTE) by linear interpolation between a datapoint and a random point among the datapoints’ $m$ nearest neighbors for $k$ times for improving data diversity and train the model on the synthesis dataset for unfairness mitigation.

Another fold of methods use sampling strategies to train a more fair model. [64] uses an adaptive sampling algorithm that evaluates the worst-performing group after each iteration using a held-out dataset and augments a random batch from this group. Besides, for each iteration, they sample a random batch from the origin dataset / the augment dataset with a probability $\lambda$ for maintaining a certain proportion of the original data in the training set. Also, [65] introduces a null-sampling procedure that can produce invariant representations in the data domain. They apply this procedure on a cVAE model and cFlow model and show a lower difference of demographic parity while having comparable performance compared with baseline models.

Besides, for task that not all the training data have group labels, researchers in [66] first train a group classifier to generate pseudo group label, by setting the classification threshold based on uncertainty. Then, they use existing in-processing method to train a fair model. This method can solve the extreme lack of sensitive attributes in medical datasets due to privacy protection.

2) In-processing method: Most of unfairness mitigation methods are applied in training procedure, which consists of designed loss function [24], [25], [67]–[73], adversarial network architecture [41], [65], [74]–[76], feature distillation [77], reinforcement learning [35]. Besides, [78] finds that CNN contains more unfairness in deeper layers, and they proposed a conservative approach that implements hierarchical features captured along the multiple layers and orthogonal regularization.

3) Post-processing method: The only research on post-processing unfairness mitigation uses a generator to generate perturbation on input image to avoid fairness-related features been extracted by the modified classification network, which allows fairness mitigation without network retraining [79].

IV. FAIRNESS IN MEDICAL IMAGE ANALYSIS

In this section, we review studies that have addressed fairness issues in training deep learning models for medical
image analysis. We first list medical datasets than have several sensitive attributes. Then, we review current researches of fairness in medical image analysis. We use the same categorization as in the previous section, including fairness evaluation and unfairness mitigation using pre-processing, in-processing, post-processing methods.

### A. Datasets

According to the definition of fairness, sensitive attributes are indispensable for analyzing whether an algorithm is fair or not. However, unlike tasks in facial recognition and NLP, there is a privacy concern among patients, which makes us harder to address the fairness issues in medical image analysis. In this part, we first survey the datasets appeared in top conference / journals in recent years. According to the type of task, we categorize the datasets with sensitive attributes into classification, segmentation and detection. The conferences / journals we choose to survey includes: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), The IEEE International Symposium on Biomedical Imaging (ISBI), IEEE Transactions on Medical Imaging (TMI) and Medical Image Analysis (MIA) and the timeline is from 2018 until present. We use task as keywords for paper selection and the datasets with sensitive attributes is listed in Tables III, IV, and V.

### B. Fairness evaluation

The foundation of unfairness mitigation is to evaluate whether there exists unfairness in a specific task. Larrazaba et al. [115] are the first to shed light on the importance of gender balance in medical image analysis. They train three deep learning models on two well-known public available X-ray image datasets for various thoracic disease diagnosis. They find that when training a DenseNet using data from one gender group, the testing performance on another gender group is much lower than that on the same gender group. The result shows that when a minimum balance is not fulfilled, the performance for unprivileged gender group deceases consistently. This finding illustrates that more attention should be paid on gender imbalance when proposing novel algorithms. On the contrary, Petersen et al. [116] train a Logistic Regression model and a 3D CNN model on ADNI dataset and analyze whether there is a linear relationship between network performance and female ratio in the training set in the AD / HC classification and sMCI / pMCI classification tasks. The result shows that there is only a weak dependence of classifier performance for male and female test subjects on the sex composition of training dataset, which to some extent proves the difference between data imbalance and algorithm fairness.

Glocker et al. [117] hold the opinion that if the test set is not balanced sampled, the bias and unfairness will remain. Besides, they use a multi-task architecture to predict sex / race and illness simultaneously and view the feature vector in latent space and find that race and sex are not encoded in their classification network.

Forde et al. [118] also find that, even the models are trained with the same training procedure (same data and same structure), there still exists significant differences in their group performances. They train a baseline X-ray classifier CheXNet on ChestX-Ray8 dataset for fifty times with the same optimizer and hyper-parameters, and find that even there

| Paper               | Conference | Sensitive Attributes | Type of Method | Description                                      |
|---------------------|------------|----------------------|----------------|--------------------------------------------------|
| Quadrianto et al.   | CVPR       | Gender               | In-processing  | Loss Function                                    |
| Wang et al. [60]    | CVPR       | Gender               | Only Evaluation| /                                                 |
| Wang and Deng [35]  | CVPR       | Race                 | In-processing  | Reinforcement Learning                           |
| Xu et al. [68]      | CVPR       | Race                 | In-processing  | Loss Function                                    |
| Ramaswamy et al. [63]| CVPR       | Age,Gender           | Pre-processing | Adversarial Network, Data Augmentation          |
| Jung et al. [77]    | CVPR       | Age,Gender,Race      | In-processing  | Feature Distillation                             |
| Tartaglione et al.  | CVPR       | Age,Gender           | In-processing  | Loss Function                                    |
| Gong et al. [69]    | CVPR       | Race,Gender          | In-processing  | Loss Function                                    |
| Liu et al. [70]     | CVPR       | Race                 | In-processing  | Loss Function                                    |
| Jeon et al. [78]    | CVPR       | Gender,Race          | In-processing  | Hierarchical Feature Extraction                  |
| Wang et al. [79]    | CVPR       | Gender,Race          | Post-processing| Perturbation Generation                         |
| Park et al. [71]    | CVPR       | Age,Gender           | In-processing  | Loss Function                                    |
| Zietlow et al. [64] | CVPR       | Gender               | Pre-processing | Data Augmentation, Sampling Strategy            |
| Jung et al. [66]    | CVPR       | Age, Gender, Race    | Pre-processing | Pseudo Group-Unlabeled Data                      |
| Wang et al. [74]    | ICCV       | Gender               | In-processing  | Adversarial Network                              |
| Wang et al. [80]    | ICCV       | Race                 | In-processing  | Loss Function                                    |
| Zhu et al. [76]     | ICCV       | Age,Gender           | In-processing  | Adversarial Network                              |
| Chen and Joo [24]   | ICCV       | Age,Gender           | In-processing  | Loss Function                                    |
| Dhar et al. [41]    | ICCV       | Gender,Race          | In-processing  | Adversarial Network                              |
| Zhao et al. [56]    | ICCV       | Age,Gender,Race      | Only Evaluation| /                                                 |
| Gong et al. [75]    | ECCV       | Age,Gender,Race      | In-processing  | Adversarial Network                              |
| Balakrishnan et al. | ECCV       | Gender               | Only Evaluation| Data Augmentation                                |
| Keilhauer et al.    | ECCV       | Gender               | In-processing  | Adversarial Network, Sampling Strategy          |
| Lokhandeet al. [73] | ECCV       | Age,Gender           | In-processing  | Loss Function                                    |

TABLE II: Fairness issues in facial recognition.
| Dataset Name           | Modality   | Body Part | Age | Gender | Racial | Marital | Height | Weight | Handedness |
|------------------------|------------|-----------|-----|--------|--------|---------|--------|--------|------------|
| PAD-UFES-20 [81]       | Dermoscope | Skin      | ✓   | ✓      | ✓      |         |        |        |            |
| ISIC [82]              | Dermoscope | Skin      | ✓   | ✓      | ✓      |         |        |        |            |
| CheXpert [83]          | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       |        |        |            |
| NIH ChestXray [84]     | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       |        |        |            |
| MIMIC-CXR [85]         | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      |            |
| PadChest [86]          | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      |            |
| BrixIA [87]            | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      |            |
| JSRT [88]              | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      |            |
| COVID-Chestxray [89]   | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      |            |
| COVID-19-CT [90]       | CT         | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      |            |
| OCTANGO [100]          | OCT        | Eyes      | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      |            |
| OCT_Manual_Delineations [101] | OCT | Eyes | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Sunnybrook [102]       | MRI        | Heart     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| ACDC [92]              | MRI        | Heart     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| UK-Biobank [103]       | MRI        | Heart     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| KiTS [104]             | CT         | Kidney    | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| BraTS [105]            | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| CANDI [106]            | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| TCGA-GBM [107]         | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| TCGA-LGG [107]         | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| Cambridge_Buckner [96] | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| ITKTube [108]          | MRI        | Vessel    | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |
| HN5SCC-3DCRT-RT [109]  | CT         | Bone      | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓          |

* https://odir2019.grand-challenge.org/

TABLE III: Medical image classification datasets with sensitive attributes.

| Dataset Name           | Modality   | Body Part | Age | Gender | Racial | Marital | Height | Weight | BMI | Handedness |
|------------------------|------------|-----------|-----|--------|--------|---------|--------|--------|-----|------------|
| ISIC [82]              | Dermoscope | Skin      | ✓   | ✓      | ✓      |         |        |        |     |            |
| COVID-Chestxray [89]   | X-ray      | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| NSCLC [98]             | CT         | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| Montgomery County X-ray [91] | X-ray | Chest | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| COVID-19-CT [90]       | CT         | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| OCTANGO [100]          | OCT        | Eyes      | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| OCT_Manual_Delineations [101] | OCT | Eyes | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Sunnybrook [102]       | MRI        | Heart     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| ACDC [92]              | MRI        | Heart     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| UK-Biobank [103]       | MRI        | Heart     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| KiTS [104]             | CT         | Kidney    | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| BraTS [105]            | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| CANDI [106]            | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| TCGA-GBM [107]         | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| TCGA-LGG [107]         | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| Cambridge_Buckner [96] | MRI        | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| ITKTube [108]          | MRI        | Vessel    | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| HN5SCC-3DCRT-RT [109]  | CT         | Bone      | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |

* https://nda.nih.gov/oai

TABLE IV: Medical image segmentation datasets with sensitive attributes.

| Dataset Name           | Modality | Body Part | Age | Gender | Racial | Marital | Height | Weight | BMI | Handedness |
|------------------------|----------|-----------|-----|--------|--------|---------|--------|--------|-----|------------|
| CheXpert [83]          | X-ray    | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| MIMIC-CXR [85]         | X-ray    | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| NIH-NLST [110]         | X-ray,Ct | Chest     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| LNDs [111]             | CT       | Lung      | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   |            |
| SunnyBrook [102]       | MRI      | Heart     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| KiTS [104]             | CT       | Kidney    | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| ADNI [112]             | MRI      | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| BraTS [105]            | MRI      | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| Cam-CAN [113]          | MRI/MRI  | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| TCGA-GBM [107]         | MRI      | Brain     | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |
| OAI* [114]             | CT       | Bone      | ✓   | ✓      | ✓      | ✓       | ✓      | ✓      | ✓   | ✓          |

* https://nda.nih.gov/oai

TABLE V: Medical image detection datasets with sensitive attributes.
is little difference between overall performance on the whole dataset among these models, the disparity of TPR between male group and female varies a lot. This result shows that unfairness may occur in the procedure of model selection.

Kinyanjui et al. [119] try to evaluate unfairness between the groups with different skin tones. They first use a segmentation model to split the dermatology image into skin part and lesion part, and measure the Individual Typology Angle (ITA) on skin part, which can represent different skin tones. Their result illustrates that there exist significant differences on classification accuracy among different skin tone groups on the SD-136 dataset.

Unlike former work, Lu et al. [120] use several epistemic uncertainty measurements to assess the disparity between different races on the Digital Mammographic Imaging Screening Trial (DMIST) dataset and find that there exists disparity on all these measurements.

Recently, Zhang et al. [121] benchmark nine different unfairness mitigation strategies on two commonly used chest X-ray datasets, CheXpert and MIMIC-CXR, and find that there is a tradeoff between the group fairness and overall classification performance. The methods that achieve group fairness usually have worse average performance for whole dataset. They also advocate for investigating unfairness-inducing mechanisms in the underlying data distribution.

C. Unfairness mitigation using Pre-processing methods

Methods including dataset resampling, dataset mix up and data augmentation are used in the procedure of pre-processing. Puyol et al. [122] mitigate unfairness on different gender and race in a cardiac MR segmentation task. By using stratified batch sampling and dataset balancing method, they decrease the value of fairness criterions a lot comparing with baseline model (nn-UNet). This work also highlights the concerning issue of unfairness in medical segmentation tasks, while all other researches focus on classification task.

Seyyed-Kalantari et al. [123] train a DenseNet-121 on three chest x-ray datasets, ChestX-Ray8, CheXpert and MIMIC-CXR separately, and find that there exists unfairness on different age and gender groups by computing TPR disparity and TPR disparity in proportion to membership. Then, they notice that when training the models using a combination of these three datasets, the disparity among groups decreases significantly, demonstrating the potential of mitigating unfairness by expanding the scale of train set.

Joshi et al. [124] believe that unfairness comes from distribution disparity on different groups, therefore, they use StyleGAN to synthesis fudus images with age-related macular degeneration (AMD) for patients with different races. This data synthesis method helps mitigate the disparity of accuracy between Caucasians and African Americans from 17.53% to 5.84%, which shows the ability of this approach for improving fairness.

D. Unfairness mitigation using In-processing methods

Current researches using in-processing methods in medical image analysis mainly consist of two directions. The first direction is adding an adversarial part on traditional baseline model, which attempts to predict sensitive attributes from input image data. An example of this kind of network architecture is shown in Fig. 6. Zhao et al. [125] train three models for HIV diagnosis from MRIs, sex prediction from adolescent brain MRIs and bone-age prediction from hand X-ray images, respectively. Sensitive attributes in consideration are age, puberty developmental score (PDS), and gender, respectively. For these three binary classification tasks, they use binary cross-entropy score as prediction loss and back-propagate the loss of confounder prediction branch with a negative coefficient. By adding this branch, the model is optimized to not distinguish sensitive attributes, and the disparity between groups is mitigated. In [126], [127], a similar method is used for mitigating unfairness on HIV diagnosis task and skin lesion diagnosis task. The major difference lies in the choice of loss functions for sensitive attribute prediction.

Li et al. [128] also propose a novel adversarial network for unfairness mitigation on skin lesion diagnosis task. Unlike former work, the structure of their sensitive attributes prediction branch is more complex. In addition to a discrimination module that predicts sensitive attributes, they also propose a critical module that predicts fairness score of the last input batch at training step. This module helps mitigate unfairness on sensitive attributes including age, gender, and skin tone, leaping a step forward comparing with discrimination-only method.

Another routine of in-processing methods focuses on the choice of loss functions, in other words, adding specific constraints or regularizers that are related with fairness criterion. Sarhan et al. [37] use a multi-task-like framework that enforces the target representation to be agnostic to sensitive attribute by maximizing the entropy between target representation for a target task and residual representation. The experiments on Heritage Health Dataset and ABIDE dataset show the decrease of accuracy disparity between different age groups and gender groups. The result of t-SNE visualization of learned embedding also proves that all sensitive attribute is mixed and a higher degree of fairness is achieved. Besides, Du et al. [129] propose a disentanglement contrastive learning method to mitigate unfairness in dermatology classification task, by using a sensitive attribute branch discarding the skin-type information and a contrastive branch improving the quality of representation. Similarly, Pakzad et al. [36] first adopt StarGAN to transform skin type of lesion images. By utilizing a regularization loss between latent representations that are extracted from images with the same diagnosis but different skin types, their model tries to learn invariant representations that can mitigate unfairness on skin tone attribute. Experiments in [130] show that, by adding fair penalties on optimization objective, the unfairness between different gender groups drops slightly. However, this type of method may lead to overparametrization and overfitting problems, which causes a fluid decision boundary that is proved to fairness gerrymandering.

There also exists other in-processing methods. Puyol-Anton et al. [122] first separate the whole train set into subsets using sensitive attributes, and train group-specific segmentation net-
work for each group. However, although the degree of unfairness mitigation is much larger than other methods at test step, it requires the accessibility of sensitive attributes at test step, which is unrealistic in practical medical applications. Therefore, this method can only be achievable for few situations and can not be generalized easily. Besides, Fan et al. [131] consider the privacy and fairness preserving simultaneously, and find that, although it cannot have considerable fairness performance with training group-specific models for each sensitive group, training with swarm learning can get more fair results than vanilla method.

E. Unfairness mitigation using Post-processing methods

Zhou et al. [132] propose an unfairness mitigation method via post-processing by integrating the prediction of PENet that takes CT image as input, and the prediction of ElasticNet that processes EHR data (contains patient medical records including demographics, vitals, inpatient/outpatient medications, ICD codes and lab test results) using a mean-pooling strategy. They find that by this post-processing, the accuracy of pulmonary embolism detection is improved significantly while keeping a relatively small disparity between different demographic groups.

Wu et al. [133] compute the saliency of each feature in the network, and use a prune strategy to remove features associated with specific group, and thus prevent to encode sensitive information into the network.

V. Experiments

In this section, we present our experiments on three medical datasets about fairness related issues, in which we explore the difference between data imbalance and unfairness, the effect of data augmentation on fairness, the existence of fairness issues in common medical image analysis tasks, and the effectiveness of some fairness mitigation methods in other areas on medical applications. By implement these experiments, our goal is to attract the attention of medical image analysis society on the fairness issue and the importance and necessity of taking fairness into consideration when developing new algorithms.

A. Difference between data imbalance and unfairness

In this section, we attempt to prove that unfairness is not equivalent to data imbalance. Our data is extracted from the open 7-classes skin lesion analysis dataset, Skin ISIC 2018 [135]. In order to access sensitive information in this dataset, we only use the training set. This dataset contains 10015 RGB images, each image is labeled as one type of seven skin lesions, including actinic keratosis (AKIEC), basal cell carcinoma (BCC), dermatofibroma (DF), melanoma (MEL), nevus (NV), pigmented benign keratosis (BKL), and vascular lesions (VASC). The original dataset is split into training (8025 images) and test (1990 images) sets with a ratio of 8:2. We use sex as the sensitive attribute. According to [127], [128], female is unprivileged group and male is privileged group. DenseNet-201 with ImageNet pretrained weights is used as classification network. Then, 7 sets of experiments with varying female: male ratios (2:8, 3:7, 4:6, 5:5, 6:4, 7:3, and 8:2) are used to analyze the relationship between fairness and data imbalance. In these experiments, we only use simple data augmentation strategies including horizontal flip, vertical flip, rotation, random resized crop and normalization in order to eliminate the interference of irrelevant factors. Besides, in order to get a comparable result with state-of-the-art algorithms, we use weighted cross entropy loss to balance the difference of the number of samples with different lesion.

The results are shown in Table VII. All the experiments are repeated for 3 times with different random seeds for avoiding random noise in train procedure. From the chart we can find that, for all fairness criteria except $\Delta$ FPR on sex attribute, the sign of criterion remains the same despite the different female: male ratios. As for $\Delta$ FPR, most of its values are positive, showing a consistence. The reason for the negative criterion on female: male ratios (2:8, 3:7, 4:6, 5:5, 6:4, 7:3, and 8:2) could be the extreme skew of the train set.

Thus, despite the distribution of different sensitive attributes in train set, the privileged and unprivileged groups remain the same, which means that unfairness arises from not only the distribution of training set, but also the inherent features of data. Another interesting phenomenon is that, even when the train set is totally balanced (Setting 4), the disparity of fairness criterion between different demographic groups is not equal to zero (even not the smallest among these 7 experiments), which is also a forceful evidence that unfairness is not the same with data imbalance.

B. Fairness in medical image classification

For classification task, several previous researches have shown that significant unfairness occurs in melanoma classification [119], [127], [128], [131], [133], chest x-ray classification [115], [118], [123], [130], [134], AMD diagnosis [124], and HIV diagnosis [125], [126]. These works illustrate that fairness does exist in medical classification tasks, and it has to be settled urgently.

In this paper, following the experiment settings in the preceding section, we also use Skin ISIC 2018 as dataset. DenseNet-201 with ImageNet pretrained weights is used for transfer learning. We split the whole dataset into train, valid
disparity is about 0.012 on that there is a little unfairness on age attribute (maximum fairness criterion are shown in Table. VIIIa, which illustrates for 3 times for averaging. The results of group prediction and test with a ratio of 7 : 1 : 2, and repeat the train step and with a ratio of 7 : 1 : 2, and repeat the train step for 3 times for averaging. The results of group prediction and fairness criterion are shown in Table. VIIIa, which illustrates that there is a little unfairness on age attribute (maximum disparity is about 0.012 on ΔFPR), while for sex attribute, it is clear that the degree of unfairness is much higher than that on age attribute (minimum disparity is about 0.012 on ΔACC and ΔNPV).

C. Fairness in medical image segmentation

There is only one research about fairness evaluation and mitigation in medical segmentation task [122], which considers the unfairness on gender and race in a cardiac MR image segmentation task. In order to evaluate whether unfairness is a common issue in medical segmentation, we choose two medical segmentation datasets with attributes from Table. IV. For each dataset, we use a state-of-the-art method for image segmentation and evaluate fairness metrics on all sensitive groups.

In this experiment, we adopt Multi-modal Brain Tumor Segmentation Challenge 2019 (BraTS19) [105], [136], [137] as dataset. BraTS 2019 datasets contain multi-institutional pre-operative MRI scans in 4 different modalities including t1, t2, t1ce and flair, and focus on the segmentation of intrinsically heterogeneous (in appearance, shape, and histology) brain tumors. It also contains age information in metadata file. Because we can only access the ground truth of training set, we split the original training set into fair_train and fair_test sets with a ratio of 7 : 3.

The nnU-Net [138] method is one of the most famous benchmark algorithms in medical image segmentation. It is the first segmentation method that is designed to deal with the heterogeneous (in appearance, shape, and histology) brain tumors. It also contains age information in metadata file. Because we can only access the ground truth of training set, we split the original training set into fair_train and fair_test sets with a ratio of 7 : 3.

and test with a ratio of 7 : 1 : 2, and repeat the train step for 3 times for averaging. The results of group prediction and fairness criterion are shown in Table. VIIIa, which illustrates that there is a little unfairness on age attribute (maximum disparity is about 0.012 on ΔFPR), while for sex attribute, it is clear that the degree of unfairness is much higher than that on age attribute (minimum disparity is about 0.012 on ΔACC and ΔNPV).

C. Fairness in medical image segmentation

There is only one research about fairness evaluation and mitigation in medical segmentation task [122], which considers the unfairness on gender and race in a cardiac MR image segmentation task. In order to evaluate whether unfairness is a common issue in medical segmentation, we choose two medical segmentation datasets with attributes from Table. IV. For each dataset, we use a state-of-the-art method for image segmentation and evaluate fairness metrics on all sensitive groups.

In this experiment, we adopt Multi-modal Brain Tumor Segmentation Challenge 2019 (BraTS19) [105], [136], [137] as dataset. BraTS 2019 datasets contain multi-institutional pre-operative MRI scans in 4 different modalities including t1, t2, t1ce and flair, and focus on the segmentation of intrinsically heterogeneous (in appearance, shape, and histology) brain tumors. It also contains age information in metadata file. Because we can only access the ground truth of training set, we split the original training set into fair_train and fair_test sets with a ratio of 7 : 3.

The nnU-Net [138] method is one of the most famous benchmark algorithms in medical image segmentation. It is the first segmentation method that is designed to deal with the dataset diversity found in the domain. It condenses and auto-mates the keys decisions for designing a successful segmen-
Table VIII: Fairness in classification. (a) Experiments are repeated for three times and results are shown by Mean ± Std. (b) Disparity of fairness metrics between demographic groups. Lower absolute value indicates better fairness. Value that lower than 0 are marked in red.

### Table VIII: Fairness in classification.

| Subgroup  | ACC ↑   | FPR ↓   | FNR ↑   | PPV ↑   | NPV ↑   | F1 ↑   |
|-----------|---------|---------|---------|---------|---------|---------|
| Age       |         |         |         |         |         |         |
| young     | 0.922 ± 0.006 | 0.157 ± 0.020 | 0.111 ± 0.007 | 0.898 ± 0.002 | 0.773 ± 0.023 | 0.892 ± 0.004 |
| old       | 0.901 ± 0.005 | 0.070 ± 0.007 | 0.233 ± 0.002 | 0.769 ± 0.003 | 0.928 ± 0.004 | 0.767 ± 0.002 |
| Sex       |         |         |         |         |         |         |
| female    | 0.924 ± 0.006 | 0.107 ± 0.009 | 0.132 ± 0.016 | 0.872 ± 0.013 | 0.871 ± 0.015 | 0.869 ± 0.015 |
| male      | 0.911 ± 0.006 | 0.088 ± 0.013 | 0.174 ± 0.008 | 0.832 ± 0.011 | 0.886 ± 0.005 | 0.828 ± 0.009 |

### D. Fairness in medical image detection

Since there is no research on fairness issues on medical detection tasks, in this section we evaluate three state-of-the-art medical detection algorithms including AlignShift [139], A3D [140], and SATr [141], which is proposed at MICCAI 2020, MICCAI 2021 and will be published at MICCAI 2022, respectively. We use Deep Lesion Dataset [33], a diverse large-scale lesion dataset proposed at CVPR 2018, which contains bounding boxes and size measurements of over 32K lesions and is widely used in medical detection tasks. We separate the dataset into train, validation and test using the original split and test using the original split in their repositories. According to [139], we use False Positive Per Image as fairness metrics: \( FFP1 = x, x \in \{0.5, 1, 2, 4, 8, 16\} \) as performance metrics, and evaluate the performance of these three methods on all samples, thin samples (thickness ≤ 2) and thick samples (thickness = 5), respectively. We use sex as sensitive attribute and split the test set into two subgroups including male and female.

The results are shown in Table. X. It is clear that for all these three methods, the performance on female group is lower than that on male group (the \( \Delta \) FFP1=16 on Thick samples using A3D and Thin samples using AlignShift is close to zero). This consistent tendency illustrates the existence of unfairness on medical detection tasks, especially in state-of-the-art algorithms. Besides, although SATr has the highest performance, it is also the most unfair methods among these three methods. This phenomenon awakes us to pay attention to the fairness of algorithm except the performance.
TABLE IX: Average performance of nnU-Net on BraTS 2019 Fair_Test Set. Dice\(_{WT}\), Dice\(_{TC}\), Dice\(_{ET}\) represents Dice Score in Whole Tumor (WH), Tumor Core (TC) and Enhanced Tumor (ET), respectively. HD\(_{95\ WT}\), HD\(_{95\ TC}\), HD\(_{95\ ET}\) represents HD\(_{95}\) in Whole Tumor (WH), Tumor Core (TC) and Enhanced Tumor (ET), respectively.

| Metrics | Age | Dice\(_{WT}\) | Dice\(_{TC}\) | Dice\(_{ET}\) | HD\(_{95\ WT}\) | HD\(_{95\ TC}\) | HD\(_{95\ ET}\) |
|---------|-----|----------------|----------------|----------------|----------------|----------------|----------------|
| Avg     | 59.92 | 0.9252         | 0.9127         | 0.8647         | 3.932          | 3.312          | 2.657          |

TABLE X: Fairness in medical image detection. SATr, A3D and AlignShift are three state-of-the-art lesion detection algorithms. (a) Group prediction performances of three algorithms on All samples, Thin samples (thickness ≤ 2) and Thick samples (thickness = 5). (b) Disparity of fairness metrics between demographic groups. Lower absolute value indicates better fairness. Values that lower than 0 are marked in red.

| (a) Group prediction performances. | FFPI ↑ | 0.5 | 1 | 2 | 4 | 8 | 16 |
|-----------------------------------|--------|-----|---|---|---|---|----|
| SATr [141]                        |        |     |   |   |   |   |    |
| All                               | 0.789  | 0.847| 0.897| 0.923| 0.948| 0.966|    |
| male                              | 0.824  | 0.880| 0.914| 0.939| 0.962| 0.973|    |
| female                            | 0.811  | 0.867| 0.907| 0.931| 0.951| 0.964|    |
| Thin                              | 0.846  | 0.892| 0.922| 0.946| 0.965| 0.973|    |
| male                              | 0.766  | 0.829| 0.885| 0.916| 0.944| 0.967|    |
| female                            | 0.811  | 0.870| 0.905| 0.935| 0.960| 0.973|    |
| Thick                             | 0.765  | 0.835| 0.881| 0.918| 0.943| 0.965|    |
| male                              | 0.784  | 0.844| 0.895| 0.931| 0.952| 0.969|    |
| female                            | 0.783  | 0.846| 0.889| 0.917| 0.947| 0.966|    |
| Thin                              | 0.799  | 0.855| 0.905| 0.936| 0.955| 0.974|    |
| male                              | 0.743  | 0.823| 0.873| 0.913| 0.939| 0.964|    |
| female                            | 0.769  | 0.840| 0.885| 0.927| 0.950| 0.963|    |
| A3D [140]                         |        |     |   |   |   |   |    |
| All                               | 0.760  | 0.825| 0.872| 0.899| 0.927| 0.950|    |
| male                              | 0.783  | 0.842| 0.879| 0.914| 0.935| 0.955|    |
| female                            | 0.791  | 0.840| 0.884| 0.910| 0.939| 0.958|    |
| Thin                              | 0.796  | 0.846| 0.890| 0.918| 0.939| 0.957|    |
| male                              | 0.731  | 0.802| 0.862| 0.888| 0.915| 0.941|    |
| female                            | 0.771  | 0.839| 0.870| 0.909| 0.930| 0.952|    |
| AlignShift [139]                  |        |     |   |   |   |   |    |
| All                               | 0.760  | 0.825| 0.872| 0.899| 0.927| 0.950|    |
| male                              | 0.783  | 0.842| 0.879| 0.914| 0.935| 0.955|    |
| female                            | 0.791  | 0.840| 0.884| 0.910| 0.939| 0.958|    |
| Thin                              | 0.796  | 0.846| 0.890| 0.918| 0.939| 0.957|    |
| male                              | 0.731  | 0.802| 0.862| 0.888| 0.915| 0.941|    |
| female                            | 0.771  | 0.839| 0.870| 0.909| 0.930| 0.952|    |

(b) Fairness criteria in detection tasks.

E. The effectiveness of fairness mitigation methods in facial recognition on medical applications

In this section, we select three unfairness mitigation algorithms used in facial recognition tasks, including one pre-processing method [142], one in-processing method [143] and one post-processing method [144].

1) Pre-processing methods: According to [142], unfairness can be mitigated by adding additional external datasets to training set. We follow this simple strategy, and use ISIC 2018 skin lesion classification dataset to train a baseline model. Then, we add PAD-UFES-20, another skin lesion dataset which contains 2,298 skin lesion images. However, among six types of lesions, only three of them are with overlapping with ISIC 2018 dataset (BCC, NEV and MEL). Therefore, we extract samples of these illness (1,114 images) and add them to training set. We train a DenseNet-201 for skin lesion classification. All the hyper-parameters are the same as experiments in Section V-B. The pipeline is shown in the lower left of Fig. 8. We repeat baseline and pre-processing experiments for three times and the results are shown using lines in orange in Fig. 9 and Fig. 10.

As mentioned before, curve at upper (lower for metric that smaller is better), more left indicates better fairness - performance trade-off and better combination property. From the figure we can conclude that simply using external dataset in training cannot mitigate unfairness effectively. Among all six metrics, none of them notices a better fairness - performance trade-off. The reason can be two folds: Firstly, due to lack of data, only few datasets can be used for joint training. However, these datasets may contains huge domain gap due to the variance of capture devices. Secondly, different with human faces, medical images with same illness varies a lot in appearance, which makes it more difficult for model to learn a proper feature representation.

Therefore, we conclude that unfairness mitigation via pre-processing methods (extending training dataset) only works on tasks that have several available external datasets, which is rare in medical tasks.

2) In-processing methods: For in-processing method, we follow the algorithm proposed in [143]. Since facial recog-
Fig. 8: Pipeline of unfairness mitigation methods, using sex as sensitive attribute for simplification. (a) Baseline model. (b) Pre-processing method: Use additional external PAD-UFES-20 dataset to train the model. (c) In-processing method: Use a sex classifier to predict pseudo sex label. Then, test sample is fed into subgroup-specific lesion classifier according to pseudo sex label. (d) Post-processing method: Use a generator to generate perturbation image that avoid the deployed model predicting sex attribute.

...the lower right of Fig. 8.

We use the same pre-processing method as V-B and use the trained baseline model as deployed model in this method. At train stage, an U-Net like generator is used to generate perturbation image and a two-layer fully-connected network is used for distinguishing the protected attribute. At inference stage, we only use the generator to generate adversarial perturbation on test samples and use deployed model for skin lesion classification. We repeat this method for three times and the result on sex and age are shown using lines in red in Fig. 9 and Fig. 10.

These two figures illustrate that, using FAAP cannot improve fairness in this skin lesion classification task. Besides, we notice a significant drop in performance in all six metrics comparing with baseline model. The reasons can be two folds: first, by visualizing the generated perturbation image, we find that its intensity is too high that affect the diagnosis of skin lesion, which absolutely causes the drop in performance. Besides, comparing to facial images, medical image are more difficult to predict sensitive attributes directly from image data, thus the effectiveness of discriminator is doubtful. In conclusion, using post-processing method (FAAP) cannot mitigate unfairness in skin lesion classification task.

From aforementioned experiments we can conclude that, to our surprise, methods that are effective in unfairness mitigation in facial recognition area seem not useful in medical image processing tasks. This phenomenon could result from domain gap in different medical datasets (pre-processing), small amount of samples in medical dataset (in-processing), difficulties in sensitive attribute prediction (post-processing), which calls upon further researches on fairness in medical image analysis. Besides, in order to mitigate unfairness in
medical applications, more attention need to be paid on the unique characteristics of medical images.

VI. CHALLENGES

In this section, we list several challenges for fairness evaluation and mitigation in medical applications.

A. The Usage of pretrained model

It is obvious that medical models are more difficult to train comparing with natural images, and pretrained models are more precious. However, most of the unfairness mitigation methods focus on pre-processing and in-processing stages, which requires the retrain of model. How to make full use of pretrained models while preserving model fairness is a big challenge for us.

B. The limitation of data size

As mentioned before, the foundation of fairness evaluation in medical image analysis is medical datasets with sensitive attributes information. However, due to the preciousness of data, the data size of medical datasets is usually smaller comparing with facial datasets. As shown in Tab. I. Only few datasets have more than 1,000 individual samples, which leads to higher uncertainty of fairness evaluation because the number of samples in each demographic group are quiet small. Two potential solutions to this challenge including finding larger datasets with sensitive attributes and using individual fairness instead of group fairness as the criterion of algorithm fairness. For the first solution, this kind of datasets might be difficult to find because the morbidity of some illness is small thus the data is difficult to collect. For the second solution, more reliable distance measurement need to be proposed for computing individual fairness scores.

C. The deficiency of sensitive attributes

The deficiency of sensitive attributes is another challenge in medical fairness evaluation. It is clear that sensitive attributes of facial recognition datasets are easier to acquire or can be inferred directly from image (for example, gender, skin tone and the type of hair). However, for medical datasets, this information is hard to acquire. Besides, in order to protect the privacy of patients, many researchers remove sensitive attributes of patients including age and gender before publish dataset, which makes it more difficult to evaluate fairness in medical applications.

D. The process on continuous sensitive attributes

Another challenge for reaching algorithm fairness is the process on continuous sensitive attributes. In former researches, experimenters usually choose thresholds by themselves without convincing reasons. For example, splitting the dataset on age with a threshold of 60 years old. This kind of process obviously influences the criterion of fairness and disturbs the evaluation of fairness. Therefore, new fairness criterion scores need to be proposed to evaluate the fairness on continuous sensitive attributes.

ACKNOWLEDGMENT

REFERENCES

[1] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Gin- nemen, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, “A review of deep learning in medical imaging: Imaging traits, technol- ogy trends, case studies with progress highlights, and future promises,” Proceedings of the IEEE, 2021.
[2] W. Saeed and C. Omlin, “Explainable ai (xai): A systematic meta- survey of current challenges and future opportunities,” arXiv preprint arXiv:2111.06420, 2021.
[3] K. Lekadir, R. Osuala, L. Gallin, N. Lazrak, K. Kushibar, G. Tsakou, S. Aussó, L. C. Alberich, K. Marias, M. Tsiknakis et al., “Future- ai: Guiding principles and consensus recommendations for trust- worthy artificial intelligence in medical imaging,” arXiv preprint arXiv:2109.09658, 2021.
[4] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, “Pulse: Self- supervised photo upsampling via latent space exploration of generative models,” in Proceedings of the ieee/cvf conference on computer vision and pattern recognition, 2020, pp. 2437–2445.
[5] A. D’Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, “Fairness is not static: deeper understanding of long term fairness via simulation studies,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 525–534.
[6] L. F. Wightman, “Lsac national longitudinal bar passage study. lsac research report series,” 1998.
[7] O. Celma, Music Recommendation and Discovery in the Long Tail. Springer, 2010.
[8] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, and Y. Zhang, “Fairness in recommendation: A survey,” arXiv preprint arXiv:2205.13619, 2022.
[9] S. Barocas, M. Hardt, and A. Narayanan, “Fairness in machine learn- ing,” Nips tutorial, vol. 1, p. 2, 2017.
[10] C. I. Esty, “Metric learning for individual fairness,” arXiv preprint arXiv:1906.00250, 2019.
[11] R. Binns, “On the apparent conflict between individual and group fairness,” in Proceedings of the 2020 conference on fairness, account- ability, and transparency, 2020, pp. 514–524.
[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.
[13] A. Narayanan, “Translation tutorial: 21 fairness definitions and their politics,” in Proc. Conf. Fairness Accountability Transparency, New York, USA, vol. 1170, 2018, p. 3.
[14] D. Li, C. T. Lin, J. Sulam, and P. H. Yi, “Deep learning prediction of sex on chest radiographs: a potential contributor to biased algorithms,” Emergency Radiology, pp. 1–6, 2022.
[15] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, “Data imbalance in classification: Experimental evaluation,” Information Sciences, vol. 513, pp. 429–441, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519310497
[16] M. W. Dong and W. Deng, “Deep visual domain adaptation: A survey,” Neurocomputing, vol. 312, pp. 135–153, 2018.
[17] J. Addour, Y. Zhu, Q. Liu, M. J. Kusner, and R. Silva, “Causal effect inference for structured treatments,” Advances in Neural Information Processing Systems, vol. 34, 2021.
[18] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proceedings of International Conference on Computer Vision (ICCV), December 2015.
[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benen- son, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[20] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federaated machine learning: Concept and applications,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1–19, 2019.
[21] S. K. Moustaapha Cisse, “Fairness and representation learning,” Google AI Brain Team, Tech. Rep., 2019.
[22] P. Kuhns and K. Shen, “Gender discrimination in job ads: Evidence from china,” The Quarterly Journal of Economics, vol. 128, no. 1, pp. 287-336, 2013.
[23] H. Yoon, “A machine learning evaluation of the compas dataset,” in 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018, pp. 5474–5474.
Fig. 9: Fairness - Performance Trade-off curves on sex attribute using different method. For (a) to (d), the closer the curve is to the top left corner, the better trade-off between fairness and performance the method have. For (e) and (f), the closer the curve is to the bottom left corner, the better trade-off between fairness and performance the method have.

Fig. 10: Fairness - Performance Trade-off curves on age attribute using different method. For (a) to (d), the closer the curve is to the top left corner, the better trade-off between fairness and performance the method have. For (e) and (f), the closer the curve is to the bottom left corner, the better trade-off between fairness and performance the method have.
recognition,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 578–586.

[69] S. Gong, X. Liu, and A. K. Jain, “Mitigating face recognition bias via group adaptive classifier,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3414–3424.

[70] C. Liu, X. Yu, Y.-H. Tsai, M. Faraki, R. Moslemi, M. Chandraker, and H. Byun, “Learning to learn across diverse data biases in deep face recognition,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4072–4082.

[71] S. Park, J. Lee, P. Lee, S. Hwang, J. Park, and H. Byun, “Fair contrastive learning for facial attribute classification,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10390–10398.

[72] E. Tartaglione, C. A. Barbano, and M. Grangetto, “End: Entangling and disentangling deep representations for bias correction,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13508–13517.

[73] V. S. Lokhande, A. K. Akash, S. N. Ravi, and V. Singh, “Fairal: Augmented langrangian method for training fair models with little regret,” in European Conference on Computer Vision. Springer, 2020, pp. 365–381.

[74] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Rotemberg, “The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience,” in European Conference on Computer Vision. Springer, 2020, pp. 12115–12124.

[75] M. Milham, M. Mennes, D. Gutman, and M. M. Barch, “Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?” IEEE transactions on medical imaging, vol. 37, no. 11, pp. 2514–2525, 2018.

[76] A.-. consortium, “The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience,” in European Conference on Computer Vision. Springer, 2020, pp. 10379–10388.

[77] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Voelkner, et al., “Patient-centric dataset of images and metadata for identifying melanomas using clinical context,” Scientific data, vol. 8, no. 1, pp. 1–8, 2021.

[78] J. Irvin, P. Rajpurkar, M. K., Y. Yu, S. Ciurea-Ilcus, C. Chute, J. Gomez, J. Y. Noordzij, J. Li, S. A. K. Amsterdam, and P. Li, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 590–597.

[79] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[80] A. J. Pollard, S. J. Berends, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, D. R. Mark, and S. Horng, “Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports,” Scientific data, vol. 6, no. 1, pp. 1–8, 2019.

[81] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Váyá, “PADchest: A large chest x-ray image dataset with multi-label annotated reports,” PLoS one, vol. 14, no. 6, pp. 1–22, 2019.

[82] A. Borghesi and R. Maroldi, “Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression,” La radiologia medica, vol. 125, pp. 509–513, 2020.

[83] J. Shiraiishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules,” American Journal of Roentgenology, vol. 174, no. 1, pp. 71–74, 2000.

[84] J. P. Cohen, P. Morrison, L. Yao, K. Roth, T. Q. Duong, and M. Ghassemi, “Covid-19 image data collection: Prospective predictions are the future,” arXiv 2006.11988, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[85] S. Haeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, and G. Thoma, “Two public chest x-ray datasets for computer-aided screening of pulmonary diseases,” Quantitative imaging in medicine and surgery, vol. 4, no. 6, p. 475, 2014.

[86] O. Bernard, A. Lalande, C. Zott, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester et al., “Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?” IEEE transactions on medical imaging, vol. 37, no. 11, pp. 2514–2525, 2018.

[87] M. Milham, M. Mennes, D. Gutman, and M. M. Barch, “Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?” IEEE transactions on medical imaging, vol. 37, no. 11, pp. 2514–2525, 2018.

[88] Y. He, A. Carass, S. D. Solomon, S. Saiidha, P. A. Calabresi, and J. L. Prince, “Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls,” Data in brief, vol. 22, pp. 629–635, 2021.

[89] D. Schoeffel, P. Elliott, J. Green, M. Landray, J. P. Hodkinson, M. Rehman, and J. Prince, “Open access ct dataset: an open-access ct image repository of 1000+ patients with confirmed covid-19 diagnosis,” BMC Research Notes, vol. 14, no. 1, pp. 1–3, 2021.

[90] M. Dizaj, J. Novo, M. G. Penedo, and M. Ortega, “Automatic extraction of vascularity measurements using oct-a images,” Procedia Computer Science, vol. 126, pp. 273–281, 2018.

[91] B. R. Johnson, S. J. Berends, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, D. R. Mark, and S. Horng, “Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports,” Scientific data, vol. 6, no. 1, pp. 1–8, 2019.
assessment, and overall survival prediction in the braTs challenge," arXiv preprint arXiv:1811.02069, 2018.

[106] J. A. Frazier, S. M. Hodge, J. L. Breeze, A. J. Giuliano, J. E. Terry, C. M. Moore, D. N. Kennedy, M. P. Lopez-Larson, V. S. Caviness, L. J. Seidman et al., “Diagnostic and sex effects on limbic volumes in early-onset bipolar disorder and schizoaffective,” Schizophrenia bulletin, vol. 34, no. 1, pp. 37–46, 2008.

[107] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle et al., “The cancer imaging archive (tica): maintaining and operating a public information repository,” Journal of digital imaging, vol. 26, no. 6, pp. 1045–1057, 2013.

[108] T. Kwitt, N. Vasconcelos, S. Razzaque, and S. Aylward, “Localizing tumor structures in ultrasound video–a phantom study,” Medical image analysis, vol. 17, no. 7, pp. 712–722, 2013.

[109] T. Bejarano, M. De Ornelas-Couto, and I. B. Mihaylov, “Longitudinal fan-beam computed tomography dataset for head-and-neck squamous cell carcinoma patients,” Medical physics, vol. 46, no. 5, pp. 2526–2537, 2019.

[110] N. L. S. T. Team et al., “The national lung screening trial: overview and study design,” Radiology, vol. 258, no. 1, p. 243, 2011.

[111] J. Pedrosa, G. Aresta, C. Ferreira, M. Rodrigues, P. Leitão, A. S. et al., “Chexvision: Fairness gaps in deep chest x-ray classifiers,” in BIOCOMPuting 2021: Proceedings of the Pacific Symposium. World Scientific, 2020, pp. 232–243.

[112] N. Joshi and P. Burlina, “Ai fairness via domain adaptation,” arXiv preprint arXiv:2104.01109, 2021.

[113] Q. Zhao, E. Adeli, and K. M. Pohl, “Training confounder-free deep learning models for medical applications,” Nature communications, vol. 11, no. 1, pp. 1–9, 2020.

[114] 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, 2021, pp. 2513–2523.

[115] S. Rotemberg, R. Raunamann, B. E. Michels, G. Schouten, and V. Cheplygina, “Risk of training diagnostic algorithms on data with demographic bias,” in Interpretable and Annotation-Efficient Learning for Medical Image Computing. Springer, 2020, pp. 183–192.

[116] X. Li, Z. Cui, Y. Wu, L. Gu, and T. Harada, “Estimating and improving fairness with adversarial learning,” arXiv preprint arXiv:2103.04243, 2021.

[117] S. Du, B. Hers, N. Bayasi, G. Harmaneck, and R. Garbi, “Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning,” in Proceedings of the 17th European Conference on Computer Vision Workshops (ECCVW 2022), 2022.

[118] V. Cherepanova, V. Nanda, M. Golblum, J. P. Dickerson, and T. Goldstein, “Technical challenges for training fair neural networks,” arXiv preprint arXiv:2102.06764, 2021.

[119] D. Fan, Y. Wu, and X. Li, “On the fairness of swarm learning in skin lesion classification,” in Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning. Springer, 2021, pp. 120–129.

[120] Y. Zhou, S.-C. Huang, J. A. Fries, A. Youssef, T. J. Amrhein, M. Chang, I. Banerjee, D. Rubin, L. Xin, N. Shah et al., “Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ech,” arXiv preprint arXiv:2111.11665, 2021.

[121] Y. Wu, D. Zeng, X. Xu, Y. Shi, and J. Hu, “Fairprune: Achieving fairness through pruning for dermatological disease diagnosis,” arXiv preprint arXiv:2203.02110, 2022.

[122] L. Luo, D. Xu, H. Chen, T.-T. Wong, and P.-A. Heng, “Pseudo bias-balanced learning for debiased chest x-ray classification,” arXiv preprint arXiv:2203.09860, 2022.

[123] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutmann, B. Helba, A. Kalloo, K. Lioprysi, M. Marchetti et al., “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic),” arXiv preprint arXiv:1902.03368, 2019.

[124] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, “Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features,” Scientific data, vol. 4, no. 1, pp. 1–13, 2017.

[125] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest et al., “The multimodal brain tumor image segmentation benchmark (brats),” IEEE transactions on medical imaging, vol. 34, no. 10, pp. 1993–2024, 2014.

[126] F. Isemeci, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” Nature methods, vol. 18, no. 2, pp. 203–211, 2021.

[127] J. Yang, Y. He, X. Huang, J. Xu, X. Ye, G. Tao, and B. Ni, “Alignshift: bridging the gap of imaging thickness in 3d anisotropic volumes,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 562–572.

[128] J. Yang, Y. He, K. Yang, Z. Lin, H. Pfister, and B. Ni, “Asymmetric 3d context fusion for universal lesion detection,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 571–580.

[129] H. Li, L. Chen, H. Han, and S. K. Zhou, “Sattr: Slice attention with transformer for universal lesion detection,” arXiv preprint arXiv:2203.07373, 2022.

[130] J. Yu, X. Hao, H. Xie, and Y. Yu, “Fair face recognition using data balancing, enhancement and fusion,” in European Conference on Computer Vision. Springer, 2022, pp. 492–505.

[131] R. Vera-Rodriguez, B. Blazquez, A. Morales, E. Gonzalez-Sosa, J. C. Neves, and H. Proenca, “Facegender: Exploiting gender information in densen face recognition systems,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Springer, 2022, pp. 492–505.
[144] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, “Face recognition: too bias, or not too bias?” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 0–1.