

Fair Machine Learning in Healthcare: A Review

Qizhang Feng¹, Mengnan Du¹, Na Zou², and Xia Hu³

¹Department of Computer Science and Engineering, Texas A&M University

²Department of Engineering Technology and Industrial Distribution, Texas A&M University

³Department of Computer Science, Rice University

August 17, 2022

Abstract

Benefiting from the digitization of healthcare data and the development of computing power, machine learning methods are increasingly used in the healthcare domain. Fairness problems have been identified in machine learning for healthcare, resulting in an unfair allocation of limited healthcare resources or excessive health risks for certain groups. Therefore, addressing the fairness problems has recently attracted increasing attention from the healthcare community. However, the intersection of machine learning for healthcare and fairness in machine learning remains understudied. In this review, we build the bridge by exposing fairness problems, summarizing possible biases, sorting out mitigation methods and pointing out challenges along with opportunities for the future.

1 Introduction

With the rapid digitization of health data and the growth of computing power, the use of machine learning has grown rapidly in many healthcare fields in recent years, leading healthcare into a new era. Many studies have attempted to implement machine learning methods on medical images, electronic health records (EHRs), clinical notes, and various other medical data [57]. For example, machine learning algorithms have been used in medical image segmentation [82], super-resolution reconstruction [83], end-to-end feature extraction [93], or relationship classification [62]. Although machine learning models have created great opportunities and numerous progress, they have also brought significant fairness challenges in the healthcare domain [99].

Machine learning methods have been found to have fairness problems in numerous areas. One well-known example is the 1998 criminal risk assessment tool COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), which was found to have bias against African-Americans to assign a higher risk score of recommitting another crime than Caucasians with the same profile [38]. With the recent development of machine learning algorithms in healthcare, there are concerns that the same fairness problems could adversely affect ethnic minorities and other under-represented communities. For example, AI models using genetic testing were more likely to misrepresent the risk of breast cancer in black women compared to white women, even though the risk was the same for both [72]. Another example is in the area of diagnostic radiology [84]. Due to differences in insurance types, Hispanic and Black patients may lack adequate medical records. This leads to machine learning models under-diagnosing them. Therefore, it is essential to design our machine learning systems for healthcare with guidance to infuse fairness more and more into them [46].

Although fairness in machine learning and health disparities have been well studied in their respective fields, the intersection of the two is still in its infancy. To mitigate the gap, this survey aims to systematically analyze what is lacking in building a bridge between healthcare systems and fair machine learning from three perspectives: 1) identifying possible biases in different stages of designing and deploying such systems, and 2) summarizing what has been done and what can be further investigated in the mitigation process

according to a different categorization on fairness measurements, 3) pointing out the specific challenges and opportunities based on uniqueness in healthcare applications towards a fair and trustable machine learning healthcare ecosystem.

In this article, we first summarize the fairness problems that arise when machine learning models are applied to different types of healthcare data in Section 2. The following Section 3 discusses potential sources of bias in the different stages of machine learning systems when applied to healthcare systems. We then present a definition of fairness in the context of machine learning in healthcare from a distributive justice perspective and a computational perspective on the measurement of fairness in Section 4. Next we review the work that mitigate the fairness problems in healthcare machine learning models according to the different stages of healthcare machine learning life-cycle in Section 5. Finally, in Section 6 we present some of the current limitations of fairness in machine learning for healthcare and list some potentially important future research challenges.

2 Fairness Problems in Healthcare

Nowadays, machine learning methods are gradually applied in high-stake applications such as employment, criminal justice, and healthcare. The digitization of medical data collection has enabled us to collect large amounts of medical data of many types and to develop machine learning algorithms for a variety of medical tasks. In this section, we present the fairness problems uncovered based on the three major types of data used, namely medical image data, structured EHR data, and textual data. In addition, we also briefly introduce the fairness issue in other data types.

2.1 Medical Image Data

Due to the success of machine learning methods in the field of computer vision, a pioneering application of machine learning to clinical data is medical image processing. Medical image has been the major data source for machine learning in healthcare [51], and the fairness problem of machine learning in medical image receives more attention.

Many machine learning models developed on X-ray image datasets (e.g., NIH Chest-Xray14, CheXpert, MIMIC-CXR and Chest-Xray8 [90, 49, 52, 84]) are found to have fairness problems. For example, a large-scale study reveals the importance of gender balance for the training of machine learning systems for the diagnosis of X-ray images [59]. Specifically, the authors build a deep neural network on the well-known NIH Chest-Xray14 dataset and the CheXpert dataset to diagnose various chest diseases under different gender imbalance conditions and find that the minority gender group consistently performs worse than the majority gender group in terms of area under the curve (AUC) metric. Other work has found that state-of-the-art classifiers differ in true positive rates (TPR) across all datasets, all clinical tasks and all subgroups [84].

Fairness problems have also been identified in machine learning methods for many other medical image datasets and task types. For example, relying on machine learning for skin cancer screening also introduces potential racial disparities [2]. Medical image segmentation makes anatomical or pathological structures in images clearer to help clinicians make accurate diagnoses. It has been suggested that there may be fairness problems based on racial differences in cine CMR segmentation models [77].

2.2 Structured Electronic Health Records Data

Electronic health record (EHR) systems store demographic information, diagnoses, radiological images, clinical records and other results and are used for tasks such as medical concept extraction, disease inference, and clinical decision support systems. EHR data can be divided into two main categories: structured data and unstructured data [88]. Structured data include demographics, diagnoses, laboratory tests, drugs, etc. Unstructured data are mainly free text, such as clinical records written by doctors and nurses. Given the

heterogeneity of these two types of data, we present them in two different sections. In this section, we focus on fairness issues found in structured EHR data.

First, the neural network-based approaches have been found to show discrimination towards certain demographic groups. For example, one related study tests and finds race-level differences in the predictions of multiple neural network models on MIMIC-IV [68]. It is found that the Black and Hispanic groups are less likely to receive the intervention, along with a shorter average duration of intervention. Similar differences are observed among groups with different marital status or insurance types.

Second, traditional machine learning models are also found to suffer from the same fairness problem. For example, a gradient boosting model is used to predict cardiovascular disease risk based on the Stanford Translational Research Integrated Database Environment (STRIDE 8) dataset [71]. There are differences in performance between subgroups with different sensitivity attributes (e.g., gender, age and ethnicity). Similar fairness issues occur in the use of pooled cohort equations (PCE) to guide physicians on whether to prescribe cholesterol-lowering therapies to prevent atherosclerotic cardiovascular disease [75]. PCE tends to overestimate risk in certain circumstances, which may lead to different groups being at risk of being under- or over-treated.

2.3 Textual Data

Unstructured EHRs are recorded in the form of clinical narratives, including medical examinations, clinical laboratory reports, operative notes and discharge summaries. They provide a wealth of personalized information for clinical decision-making. Here we use textual data to refer to unstructured EHR data. Natural Language Processing (NLP) methods study theories and methods for effective communication between humans and computers using natural language. In recent years machine learning methods are being applied to learn embedded representations of medical concepts [20]. There is an increasing trend from the community to investigate the fairness issues of applying machine learning methods to textual data. We divide current research of identifying fairness issues into two categories: fairness issue of intermediate word or sentence representations and fairness issue of model predictions.

This first category focuses on the fairness issue of the representations (or embeddings) obtained from the textual inputs. For example, word embeddings from a large pre-trained deep embedding model (i.e. BERT) are used to identify [97]. In this work, the authors design a fill-in-the-blank task based on clinical notes from the MIMIC-III dataset. Potential relationships captured by the word embedding vector that are dangerous in terms of gender are identified by comparing the probability gap between different fill-in gender pronouns. The cosine similarity of word embeddings can also be used to measure fairness issues [70]. The authors quantify the “genderness” nature of words by calculating rank turbulence dispersion (RTD) rankings from the cosine similarity of word embeddings and gendered clusters (male and female).

The second type of approach is to identify fairness problems by directly comparing the predictions of the NLP models across demographic groups. For example, machine learning methods also reflect a pernicious tendency towards social bias in automated question answering (QA) tasks [63]. This work presents Q-Pain dataset for the medical QA task in the context of pain management. The authors evaluate the two models GPT-2 and GPT-3[78, 13] through a newly proposed potential bias framework and find statistically significant differences in treatment decisions across gender subgroups.

2.4 Other Data Types

In addition to the three types of data mentioned above, data such as physiological data and genomics data are used in healthcare applications. Several studies have also identified fairness issues that arise when machine learning methods are applied to these data. For example, the Vital-ECG, a wearable smart device that monitors ECG and chest X-ray signals, is embedded with machine learning methods to predict and monitor arterial blood pressure [73] and is found to underestimate the risk of disease in female patients. Similar issues of fairness are found in models using genetic data. A polygenic risk model trained to analyse schizophrenia

using only European data is found to have reduced performance in East Asian populations [58].

Multimodal models can simultaneously accept data input from different modalities including clinical text data, image data, waveform data and biomedical data. The multiple perspectives of medical data from different modalities provide information on patient treatment, allowing multimodal models to gradually show their unique advantages in the healthcare field [54]. The research on the fairness issue of multimodal models is not many at current stage [100, 17]. A preliminary work presents a multimodal benchmark dataset consisting of 1794 patients and their corresponding EHR data and high-resolution computed tomography (CT) data, called RadFusion [100]. The authors evaluate the performance of several representative multimodal fusion models on a diagnostic task of pulmonary embolism and benchmark their fairness in a protected subgroup. They find that the multimodal model improves diagnostic performance across the board, without introducing large differences in TPR between populations.

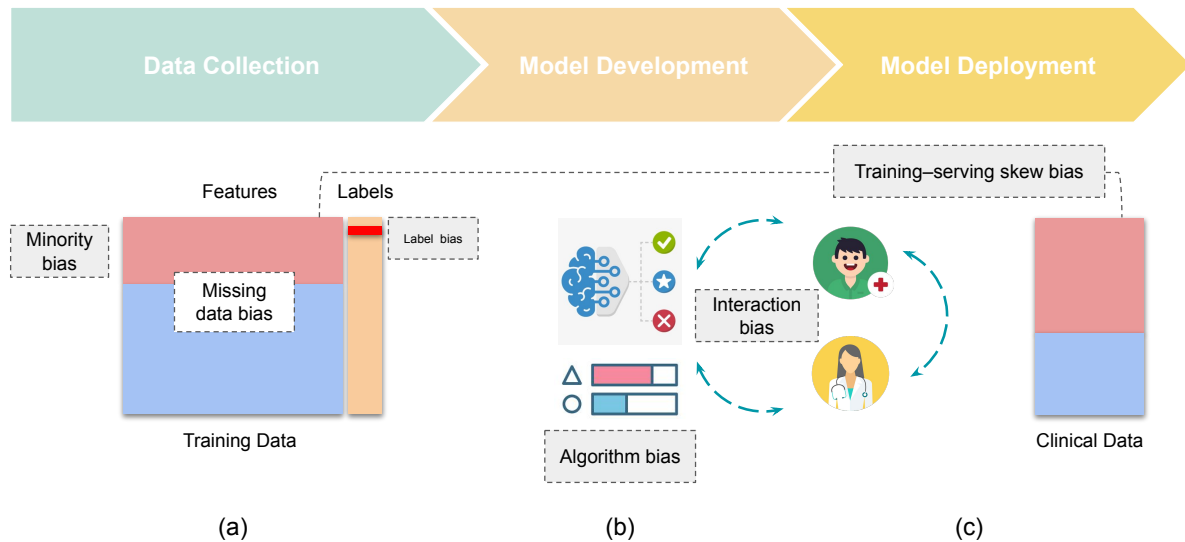


Figure 1: **Bias at the different stages in machine learning systems:** (a) The biases that exist at the data collection stage include minority bias, missing-data bias and label bias. (b) Algorithm bias that exists in model development stage can lead to unfair outcome. (c) The biases that exist at the data collection stage include interaction bias and training-serving skew bias.

3 Sources of Fairness Problems in Machine Learning

We have described in last section that many machine learning-based models suffer from fairness problems, and in this section, we attribute the reasons for these fairness problems. We use the term bias to denote the cause of the fairness problem [67], and there are various complex biases in the healthcare field. The process of building a machine learning-based healthcare system can be divided into three stages. First, the agency collects relevant clinical data for model development. Then the developers select and train a suitable model for the intended task, based on the data and the type of task. Finally, the model can be licensed by the institution concerned for deployment in real clinical practice. We introduce below the types of bias that can occur in the three phases (also refer to the overview in Figure 1).

3.1 Bias in Data Collection

Data collection is the first stage at which bias may be introduced. A machine learning model is trained to fit the distribution of the training data, and when there is bias in the data, the model perpetuates the bias. In the following paragraphs, we review several common types of data bias in clinical practice.

3.1.1 Minority Bias

Minority bias occurs when the sample size of a demographic group is smaller than that of other groups. The development of machine learning algorithms in healthcare is currently highly dependent on public biobank databases [9, 14]. However, due to the uneven development of medical standards, most of the data collection is done in Europe. This has led to human knowledge of the disease being conducted using biobank repositories representing mainly individuals of European ancestry. For example, the vast majority of cases in the Cancer Genome Atlas (TCGA) are made up of whites, accounting for approximately 82.0% of cases. In contrast, a very small proportion of cases are from Black, Asian and other ethnic minorities [39]. Indeed, demographic data such as ethnicity are crucial in determining the mutational profile and mechanisms of cancer. As a result, genetic risk models perform worse in ethnic minority populations.

3.1.2 Missing-data Bias

Missing data bias occurs when data may be missing in a non-random way. Machine learning algorithms may cause harm to people with missing data in the dataset. For example, research has found that vulnerable people who are of low socioeconomic status are likely to be seen in a piecemeal fashion, or are unable to be seen. If patients are identified based on a certain number of ICD codes, records of the same number of visits to several different healthcare systems for these patients may be missing. Another example is that despite numerous initiatives to include sexual orientation and gender identity in electronic health records, this information has been largely absent to date. Clinical decision support systems based on machine learning can misinterpret lack of access to care as a lower burden of disease and therefore produce inaccurate predictions for these groups [18].

3.1.3 Label Bias

Bias may also be present in data labels. The quality of the labels can contribute to bias [79]. For example, people of low socioeconomic status may be more likely to be seen in teaching clinics, where documentation or clinical reasoning may be less accurate or systematically different from the care provided to patients of high socioeconomic status. Algorithms based on these data may reflect practitioner bias and misclassify patients based on these factors. The choice of inappropriate labels can also introduce bias. For example, some models use specific phrases that appear in clinical records as proxy labels indicating the presence of cardiovascular disease. However, because women have different symptoms of acute coronary syndromes, proxy phrases have different meanings for men and women. As a result, women may receive delayed care, thus exhibiting a gender bias in labeling.

3.2 Bias in Model Development

Algorithm design choices, such as the use of certain optimization functions, regularization, or the use of statistically biased estimators, can lead to biased algorithmic decisions.

3.2.1 Algorithmic Bias

Algorithmic bias is where the source of bias can be traced back to the model itself, and where the model systematically leads to unfair results for certain groups. Machine learning models can perpetuate or even amplify biases in the data. This is because machine learning models aim to maximize the overall predictive performance of the training data, which may optimize for individuals that occur more frequently than others, while neglecting to learn good predictions for groups that are under-represented in the data due to sampling bias. As a result, a model may perform better overall, but perform poorly in underrepresented groups. For example, one study indicates that different machine learning algorithms can exhibit different degrees of bias on the same dataset [91]. Specifically, the authors compare the performance of three different models, namely logistic regression, random forest and XGBoost. They find that the XGBoost model had the most significant difference in AUC across all racial groupings, while the logistic regression model has the least difference. Another studies indicates that it is possible for the models to capture and amplify the association between labels and sensitive attributes even in balanced datasets [89]. The authors show that even if each label co-occurs with each gender the same number of times, the learned model would amplify the association between label and gender as if the data were not balanced.

3.3 Bias in Model Deployment

A trained machine learning model can be applied to clinical practice when it has passed regulatory authorization. The bias is likely to occur at this stage. In this section, we focus on two types of bias.

3.3.1 Training-serving Skew Bias

The training service skew bias is due to the fact that the data distribution encountered by the model in the deployment environment is different from the data distribution at the time of training. This phenomenon is known as dataset shifting. During model training, a strong assumption is that the training and test datasets are drawn independently and exactly from the same distribution (i.i.d.). This can lead to fairness issues when the model is deployed, even if it satisfies the notion of fairness in the training dataset. The phenomenon of data shifting may occur with racially skewed public biobank datasets, thereby having a differential impact on ethnic subpopulations. For example, the first AI model to exceed clinical rank in predicting lymph node metastasis was trained and evaluated on the CAMELYON16/17 dataset, which is unique to the Netherlands [9, 48]. In addition to changes in ethnicity in the population, changes in medical equipment, such as image capture and biometrics, can also lead to bias. For example, in radiology, there may be differences in radiation dose that affect the signal-to-noise ratio of the images obtained. In pathology, there is also a great deal of heterogeneity in tissue preparation, staining protocols and specific scanner camera parameters, which has been shown to affect the performance of models in cancer diagnostic tasks. Data sets may also change in response to technological developments or changes in human behavior. A typical example includes the migration of ICD-8 to ICD-9 [45]. Another example is the discontinuation of the Epic sepsis model (ESM) due to changes in patient demographics as a result of COVID-19. To date, most work has focused on short-term learning of fairness classifiers, and there has been no research on the analysis of fairness metrics under temporal dataset transfer.

3.3.2 Interaction Bias

The last type of bias arises from the interaction of the model with its users. On the one hand, protected groups may distrust a model’s predictions in light of a history of exploitation and unethical behaviour, believing that the model is biased against them. This is also referred as informed mistrust bias. On the other hand, clinicians can also place too much trust in machine learning models and inappropriately act on inaccurate predictions, which can be called automation bias [79].

4 Measurement of Fairness

Governments have introduced anti-discrimination laws to prohibit unfair treatment of individuals based on specific characteristics. These specific characteristics are also referred to as sensitive attributes, such as gender, race, etc. It is crucial to choose a proper fairness measurement for a certain scenario. In this section, we first introduce two fairness principles inspired by distributive justice, and then summarize the common measures of fairness that apply to them.

To measure the fairness of a given decision algorithm $f(\cdot)$, we define $\mathbf{x} \in \mathbb{R}^{d_x}$ as the nonsensitive features vector and $\mathbf{z} \in \mathbb{R}^{d_z}$ as the sensitive features vector. In most cases, only one sensitive feature is considered, so we use z when $d_z = 1$. The prediction of the model $f(\cdot)$ with input \mathbf{x} as $\hat{y} = f(\mathbf{x})$, and y is the corresponding ground truth label. Without loss of generality, we consider a binary classification problem where $\hat{y}, y \in \{0, 1\}$. Other symbols and definition can be found in Table. 1

4.1 Distributive Justice in Machine learning for Healthcare

Distributive justice is concerned with the distribution of resources among members of a society, and the underlying idea of distributive justice theories are distribution principles and metrics of justice. While the distribution principles define the rules on how resources should be distributed, the justice metric defines the type of resources [55]. We can understand fairness in machine learning under healthcare scenario from principles of distributive justice, where the resource is medical resources or prediction error rates. Different healthcare scenarios can require different fairness policies in the application of machine learning models to

Table 1: Main symbols and definitions.

Symbol	Definition
$f(\cdot)$	A machine learning model that maps attributes to predictions.
$\mathbf{x} \in \mathbb{R}^{d_x}$	The non-sensitive attributes with a dimension of d_x .
$\mathbf{z} \in \mathbb{R}^{d_z}$	The sensitive attribute.
z	The sensitive attribute when $d_z = 1$.
$\hat{y} \in \{0, 1\}$	A binary prediction that indicates negative and positive outcomes for 0 and 1, respectively.
$y \in \{0, 1\}$	A binary ground truth that indicates negative and positive outcomes for 0 and 1, respectively.
$\hat{y}_{\{z \leftarrow a\}}$	A prediction in counterfactual world if $z = a$.
\mathcal{D}	The training dataset.
$d(\cdot, \cdot)$	Distance of two individuals in the attribute space.
$D(\cdot, \cdot)$	Distance of two individuals in the prediction space.
θ	Parameters θ of the backbone network.
ϕ	Parameters ϕ of the adversarial network.
$A_\phi(\cdot)$	Adversarial network with parameters ϕ
$L(\mathcal{D}; \theta)$	The downstream task loss.
$L_{adv}(\mathcal{D}; \phi)$	The adversarial loss.

healthcare settings, and they can generally be categorized into two classes: equal performance, and equal allocation [79].

4.1.1 Equal Performance

Equal performance means that a model is guaranteed to be equally accurate for patients in protected and non-protected groups. Performance can include things such as equal sensitivity (also called equal opportunity [95]), equalized odds, and equal positive predictive value [21], or broader metrics such as AUC, etc. A case study can show the scenario where equal performance is needed. The machine learning system may be introduced to build a monitoring system which is used to warn a rapid steam when inpatients are at high risk for deterioration [36]. If the predictive model imposes a high false positive rate on the protected group, patients in the protected group may lose the opportunity to be identified, leading to serious consequences. However, forcing a model’s predictions to have one of the performance characteristics of equality [42] may have unintended consequences. For example, the model may achieve equal odds by sacrificing the accuracy of the unprotected group, which undermines the benefit principle. In the following, we introduce some fairness measurements that follow the principle of equal performance.

- **Equalized Odds** requires that the decision rates across demographic subgroups be the same when their outcome is the same:

$$\mathbb{P}\{\hat{y} = 1 \mid z = a, y = y_b\} = \mathbb{P}\{\hat{y} = 1 \mid z = b, y = y_a\}, \quad (1)$$

where $y_a = y_b$ indicates that the outcome is the same for the subgroup with sensitive attributes of a and b . When it is a binary classification task, Equalized Odds requires the algorithm to have equal true positive rates as well as equal false positive rates.

- **Equal Opportunity** is preferred when people care more about true positive rates. We say a classifier satisfies the equal opportunity if the true positive rate is the same across the groups [95]:

$$\mathbb{P}\{\hat{y} = 1 \mid y = 1, z = a\} = \mathbb{P}\{\hat{y} = 1 \mid y = 1, z = b\} \quad (2)$$

for subgroup a and b . It can also be referred as positive predictive value parity. Similar, there is the negative predictive value parity:

$$\mathbb{P}\{\hat{y} = 1 \mid y = 0, z = a\} = \mathbb{P}\{\hat{y} = 1 \mid y = 0, z = b\} \quad (3)$$

Predictive value parity is also called sufficiency.

- **Treatment Equality** requires that the ratio of false negatives and false positives be the same for subgroups [10]:

$$\frac{\mathbb{P}(\hat{y} = 1 | y = 0, z = a)}{\mathbb{P}(\hat{y} = 0 | y = 1, z = a)} = \frac{\mathbb{P}(\hat{y} = 1 | y = 0, z = b)}{\mathbb{P}(\hat{y} = 0 | y = 1, z = b)}. \quad (4)$$

4.1.2 Equal Allocation

Machine learning models are often used for the allocation of healthcare resources. Ensuring that resources are allocated proportionally to patients in the protected group means that they are allocated equally (also known as demographic parity). Equal allocation differs from equal performance in that allocation is determined by a positive prediction rate regardless of its accuracy. A typical scenario in which the principle of equitable allocation is applied more appropriately is the issue of vaccine distribution. When the machine learning model considers only the overall optimal prevention strategy, its allocation strategy can be detrimental to the protected group. Equal allocation is also more applicable when the label is historically biased. For example, if historically African American women have been sent for such procedures at unduly low rates, then a 'correct' prediction based on historical data would underestimate the status of these women. Thus, equal allocation can be used to correct for past biases [80]. In the following, we introduce some fairness measurement methods that follow the principle of equal allocation.

- **Demographic Parity (DP)** is satisfied if a decision-making algorithm gives equal decision rates regardless of the outcome:

$$\mathbb{P}(\hat{y} = 1 | z = a) = \mathbb{P}(\hat{y} = 1 | z = b), \quad (5)$$

for different demographic subgroups a and b . DP can be extended for multi-class classification application such as image recognition, text categorization, etc [27]:

$$\sum_{k=1}^K |\mathbb{P}(\hat{y} = k | z = a) - \mathbb{P}(\hat{y} = k | z = b)| = 0, \quad \forall k \in [K], \quad (6)$$

where $[K] = \{1, \dots, K\}$ indicates K number of classes. An alternative definition can constitute the summation with a maximum. DP can also be applied for the regression model rather than the classification model [22]:

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(\hat{y} \leq t | z = a) - \mathbb{P}(\hat{y} \leq t | z = b)| = 0. \quad (7)$$

- **Fairness through Unawareness (FTU)** [56] defines an algorithm to be FTU fair as long as sensitive attributes are not used by the decision-making algorithm $f(\cdot)$:

$$\mathbb{P}(\hat{y} | \mathbf{x}, z) = \mathbb{P}(\hat{y} | \mathbf{x}) \quad (8)$$

However FTU will fail due to "redundant coding". This means that even if a particular sensitive attribute is not present in the data, a combination of other non-sensitive features can act as a proxy.

- **Fairness through Awareness** [34] emphasizes that a fair algorithm should give similar decisions for two individuals with similar non-sensitive attributes:

$$D(f(x), f(x')) \leq d(x, x') \quad (9)$$

for two individuals, x and x' from different subgroups. And the algorithm should satisfy the (D, d) -Lipschitz property.

- **Counterfactual Fairness** [56] is derived from causal theory. The intuition of counterfactual fairness is that a fair algorithm should provide the same decision for a real-world individual and its corresponding one in the counterfactual world:

$$\mathbb{P}[\hat{y}_{\{z \leftarrow a\}} = c | \mathbf{x}, z = a] = \mathbb{P}[\hat{y}_{\{z \leftarrow b\}} = c | \mathbf{x}, z = a] \quad (10)$$

However, it is very hard to arrive at a consensus on what the causal graph should be.

5 Mitigation of Fairness Problem

We introduce the the methods to mitigate the fairness problems in the stage of data collection, model development, and model deployment.

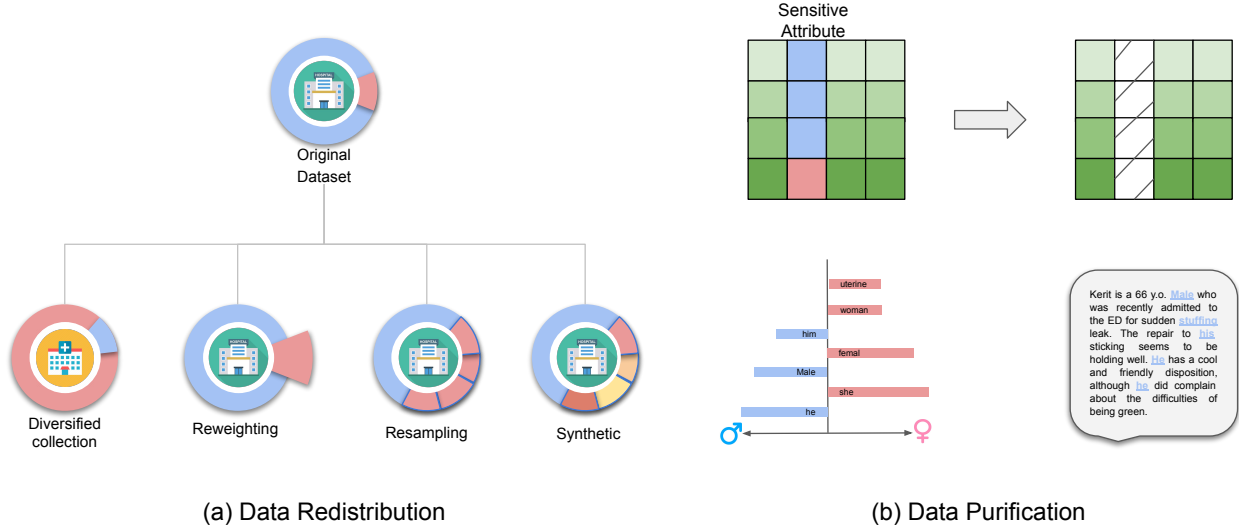


Figure 2: **An illustration of methods mitigating the fairness problem in data collection stage:** (a) Data redistribution methods adjust the distribution of the data. The diversified collection method collects data from other hospitals. The reweighting method assigns the weights to minority data. The resampling method seeks to create fair training samples in the sampling strategy. The synthetic method generates fake data. (b) Data purification methods remove sensitive information directly from the data. For example, removing sensitive attributes from tabular data or removing gender-specific pronouns from textual data.

5.1 Mitigating Fairness Problem in Data Collection

Data bias can be transferred and embedded in machine learning models. We can therefore mitigate fairness issues during the data collection phase. These methods are divided into two groups, data redistribution methods and data cleansing methods.

5.1.1 Data Redistribution

The ill-fitting data distribution is the source of fairness problem. For example, in the data sets collected in Europe, there are far fewer data from groups such as Black, Asian ethnic minorities than from whites. To alleviate this problem, the data redistribution approach attempts to adjust the distribution of the data and improve the predictive quality of underrepresented groups. A straightforward idea is to collect more comprehensive data from external sources. However, given the issues of patient privacy and data collection costs, it is often unrealistic to require direct data sharing between institutions. Federated learning is a promising technique for this purpose [18]. Instead of training a machine learning model on a single centralized dataset, federated learning is a distributed learning paradigm that allows users on a network to jointly train a common model without sharing the original dataset. For instance, Swarm Learning (SL), a federated learning variant that combines edge computing and block-chain, is evaluated on the large publicly available skin lesion dataset Skin ISIC 2018 [37]. They find that SL improves fairness performance compared to centralized

training. However, even in the case of extensive data collection, federated learning does not always guarantee balanced data. An alternative solution is to directly adjust the data distribution or weights in the original dataset. Reweighting assigns the weights to the training data to indicate the “importance” of them. The reweighting method has been used in skin lesion classification [94] or Alzheimer’s disease diagnosis [87]. The disadvantage of classifiers trained with weighted samples is that they may not have robust performance and may result in high variance of the estimator. Resampling can adjust distribution via sampling a more fair sub-samples of the original training dataset. It intuitively corrects for underrepresentation of minority sub-populations by balancing the sub-samples of the original dataset. SMOTE [16] is a combination of the method of oversampling the minority group and under-sampling the majority group, and has been used to improve survival prediction in patients with heart failure. Resampling methods to correct for underrepresentation also suffer from reduced diversity of characteristics. Beside utilizing the data from the real dataset, synthetic data can also be used to address the fairness problem [85]. The authors propose a self-supervised learning process that injects synthetic data into the training dataset. They can control the bias in the data by incorporating fairness constraints. By combining GAN and federated learning, FELICIA (FEderated LearnIng with a CentralIzed Adversary) can generate high-quality synthetic images to improve distribution coverage [81]. Synthetic data is also used to address the high cost and privacy issues of data collection, and it is crucial to ensure that synthetic data can fairly represent minority groups [11].

As the equal allocation indicates to remove the sensitive attribute information and let the machine learning model unable to discriminate the membership, the fairness through unawareness and the data purification method introduced above directly remove the sensitive information. The federated learning method, the synthetic data method, and the sampling method balance the distribution of data from various subgroups to meet the principle of equal performance.

5.1.2 Data Purification

Unfairness is inherent in our society. As such, data collected from the real world may contain bias itself. The data purification approach is guided by straightforward intuition and mitigates fairness problem by adjusting the data itself.

Fairness through unawareness removes the sensitive protected attributes from the input of the algorithm. However, denying protected attribute information has been shown to be insufficient in non-discrimination and satisfying fairness guarantees for many applications, as other input features or their combination can play as proxy variables that correlate with protected group membership.

The data purification method has also been investigated to mitigate the fairness problem in the language model for NLP tasks where the input is clinical notes [69]. Specifically, the authors quantify the “genderedness” of n-grams for both male and female patient groups based on cosine similarity between word vectors generated by the pre-trained BERT-base [28] and Clinical BERT [5] word embeddings. These words are then ranked using rank perturbation dispersion (RTD) [29]. The authors then trim the clinical notes by directly removing the most biased n-grams.

Sometimes bias is present not only in the data features but also be encoded in the outcome, which means that the labels may be biased. The data massaging method chose to change the labels of some objects in the dataset [53]. The application of data massaging in healthcare domain is yet to be explored.

5.2 Mitigating Fairness Problem in Model Development

Machine learning models may inherit, amplify or even lead to bias. We present two categories of approaches to mitigate fairness problem during the model development stage, namely model desensitization and model constraint.

5.2.1 Model Desensitization

Some works suggest that machine learning models can differentiate between sensitive information in the absence of direct sensitive information, and thus further differentiate the treatment of patients. The idea that model desensitization approaches mitigate fairness issues lies in the complete elimination of the models’

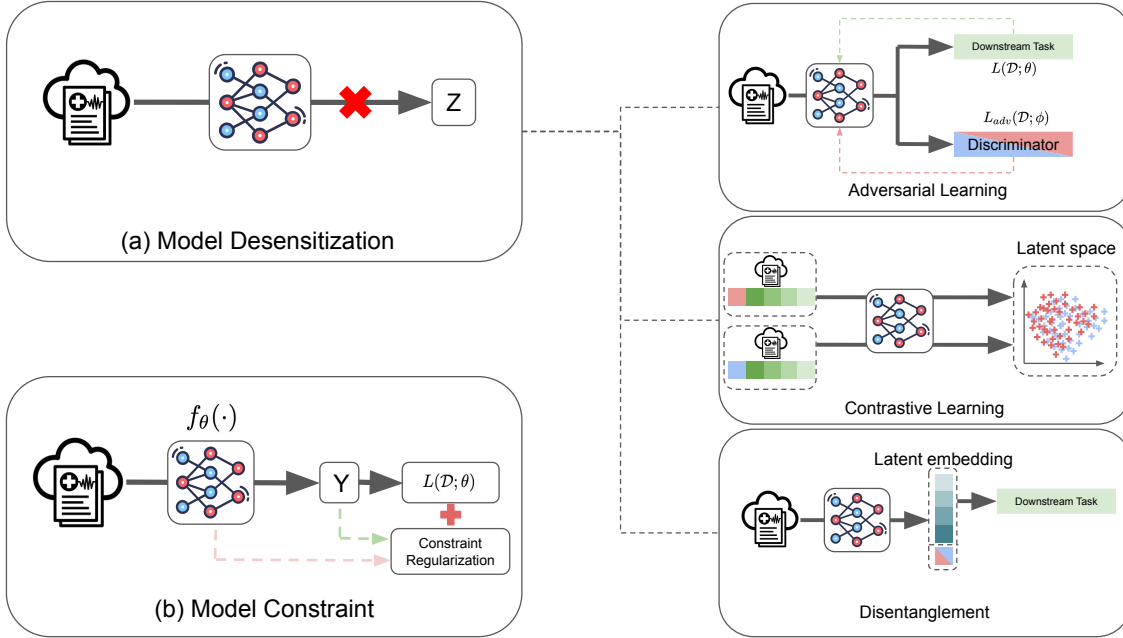


Figure 3: **An illustration of methods mitigating the fairness problem in model development stage:** (a) Model desensitization removes the ability of the model to discriminate between sensitive attribute information. Adversarial learning disables the model of predicting sensitive attributes. Disentanglement method separates and removes the sensitive attribute information from latent embedding. Contrastive learning enforces the samples with various sensitive attributes to be close in latent space. (b) Model constraint methods add additional constraints or regularization term.

ability to discriminate sensitive information.

Adversarial learning is a widely used method to debias a model. Specifically, the goal of adversarial learning is to enable the model to complete downstream tasks while failing to predict sensitive attributes. Adversarial learning is first introduced in Generative Adversarial Networks (GANs) [41] and then applied to fair machine learning [66]. Fair adversarial learning generally contains two branches, one is for downstream tasks while the other one is to remove information of sensitive attributes:

$$\min_{\theta} \max_{\phi} L(\mathcal{D}; \theta) + L_{adv}(\mathcal{D}; \phi), \quad (11)$$

where \mathcal{D} is the training dataset. θ is the parameters for the downstream task, and ϕ is the parameter for the adversarial classification. L is the normal object function, and L_{adv} is the adversarial object function that indicates the error to predict sensitive attributes. Adversarial learning is applied to debias a model for the diagnosis of chest X-ray and mammograms [23]. The authors use CNN with two branches, where one predicts the classification target, the other predicts the sensitive attributes. The training has two steps. The first step minimizes the loss for both branches. In the second one, a flipped sign gradient of adversarial branch is backpropagated, with the aim of suppressing learning of protected variables. In addition to use cases for medical image data, adversarial learning has also been used to build fair machine learning models that can handle both EHR data and textual data [75, 97].

Some other model desensitization approaches have been proposed in the context of general fairness problem instead of healthcare. The disentanglement method assumes that the entangled information from the input

space could be disentangled in the latent embedding space. To make downstream tasks fair, the disentanglement method separates and removes sensitive information in the latent embedding space. Existing work has explored the use of the Variational Autoencoder (VAE) to achieve group and subgroup fairness with respect to multiple sensitive attributes [24]. Contrastive learning method projects the input data into the latent space and encourages data points with various sensitive attributes to be close in the latent space and the data points with the same sensitive attributes to be scattered. Some work has explored the use of contrastive learning methods to debias the pre-trained text encoder [19], or to remove the effect of gender information in self-supervised embedding [86].

5.2.2 Model Constraint

Compared to model desensitization method which implicitly debias the model, model constraint method mitigates the fairness problem by directly add constraints. This is usually accomplished by adding additional optimization objectives to meet fairness requirements. This optimization objective can directly improve the fairness metric [3]. Or it can also includes regularization term to fulfill non-discrimination principle. It can satisfy the counterfactual fairness [74]. The authors develop an augmented counterfactual fairness criteria to reduce biases implicitly embedded in EHR data. The core idea is to require the machine learning model give the same prediction for a patient and a counterfactual patient after changing the sensitive attribute. The optimization object contains three components. The first two terms are the prediction losses for the factual and counterfactual samples. The last term is an additional regularization term, using counterfactual Logit Pairing (CLP) [40] to encourage the algorithm to meet the proposed criteria. The key drawbacks of the model constraint method is that the optimization object is often non-convex, which has been shown to reduce performance. Also The strength of regularization does not equally affect the fairness metric.

Since the main idea of model desensitization is to remove sensitive information and make the model non-discriminatory, model desensitization methods are more suitable for equality allocation. However, by setting appropriate optimizations, certain model desensitization methods have been shown to be able to be used for both equal allocation and equal performance [66]. In this work, the authors designed adversarial objects to constrain population parity as well as equalization odds. The model constraint approach is more flexible to satisfy equal allocation or equal performance, as the fairness constraint can be used directly as an optimization object.

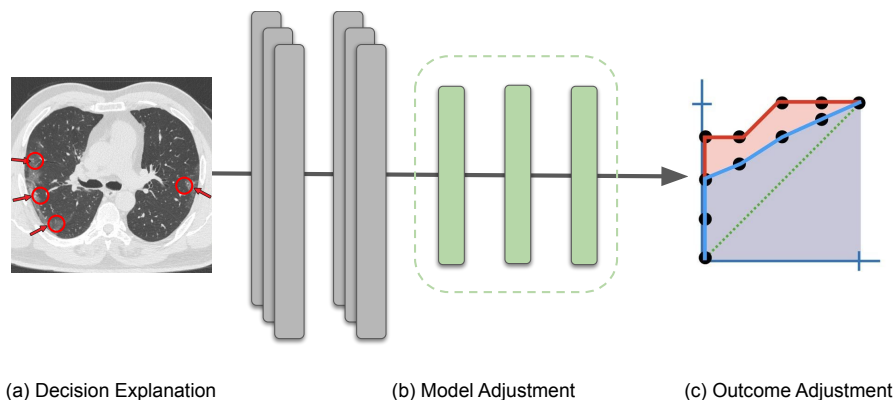


Figure 4: **An illustration of methods mitigating the fairness problem in model deployment stage:** (a) The decision explanation method offers the explanation to the outcome via XAI tool. (b) The model adjustment method fine-tunes the last few layers of the deployed mode. (c) The outcome adjustment method adjusts the original outcome to meet fairness requirement.

5.3 Mitigating Fairness Problem in Model Deployment

When a trained model is deployed into clinical practice, the bias that exists at this stage may become apparent. Given the complexity of training models, building an unbiased model from scratch is often expensive, or

even impossible. Mitigating the fairness problem during the model deployment stage can provide a flexible and effective approach to address this issue. In this section, we present three types of mitigation methods, namely, decision explanation, model adjustment, and outcome adjustment.

5.3.1 Decision Explanation

Ensuring fairness of machine learning systems is a human-in-the loop process. The decision explanation approach proposes the use of explainable artificial intelligence (XAI) methods to mitigate fairness problems from multiple dimensions at the model deployment stage.

Fairness problem can arise when users over-trust an unfair model. In high-risk scenarios, fully automated decision-making is often undesirable. Sniffing out fairness issues can tell when to trust or distrust AI [7, 32, 8, 31], which can further be combined with the domain knowledge of human experts to complement the model [98]. XAI can be used as a tool to expose biases in machine learning models via revealing the relationship between model explainability and prediction fairness [68]. The authors use several XAI methods to analyze the feature importance of the trained models. It is discovered that demographic features are important for algorithm decision and that different treatment exists for different groups of patients by race, gender, and age. They also rank the features for different subgroups and find that sensitive attributes are more important for specific groups.

On the other hand, even if the model is fair, fairness problem may arise when users have insufficient trust in the model [30], leading to the possibility that patients may avoid seeking treatment from clinicians or systems that use the model, or deliberately omit information [79]. Related studies have noted that interpretation of model decisions has been shown to increase trust in the system [1] as well as trust in medical experts [47].

Finally, the XAI approach has also been used to improve the performance of machine learning models in healthcare by embedding relevant prior knowledge into the model [92, 35]. Following a similar idea, we may also embed fairness prior knowledge into the model in the expectation of regularizing the model's behavior [33]. Fairness and explainability of machine learning models fall within a broader scope called ethical machine learning. The intersection of fairness and explainability of machine learning models is a promising area of research that will be discussed in chapter 6.4.

5.3.2 Model Adjustment

The assumption that that data in the deployment phase is i.i.d. with the data in the training phase is often not true. A naive solution is to retrain the entire model, which can be expensive and impossible. The model adjustment method seeks to solve the problem by fine-tuning part of the model. Transfer learning may provide a simple and effective way to mitigate the problem [50]. This work proposes solving the shortcut problem where the model relies on simple and shallow features (e.g. the sensitive attribute) to make the decision. Specifically, the training pipeline contains two stages. The authors first train the model on an unbiased dataset. Then the model is tuned on a new biased dataset with only the last few layers being fine-tuned. The results show that the proposed approach improves the generalization performance on aged people.

5.3.3 Outcome Adjustment

The outcome adjustment method aims to directly adjust the results of the original model. The core idea of the outcome adjustment method is to redistribute the prediction of the model to give a more favorable outcome to the protected group. The calibration proposes to ensure that the distribution of positive predictions in each subgroup is consistent with the distribution of positive samples [26]. However, calibration has been shown to be incompatible between different fairness standards when calibrating to multiple protected groups [15]. For example, calibration is in conflict with equalized odds and disparate impact [76]. And calibration is highly inappropriate for the healthcare system due to the neglect or even sacrifice of individual predictions [18]. Thresholding adjusts the decision-making boundaries for demographic groups so that the protected group can be positively classified. For example, the authors adapted the classifier to meet the equal odds or equal opportunity requirement by applying various thresholds based on sensitive attributes. [43].

6 Research Challenges

The fairness problem in machine learning for healthcare has received increasing attention in recent years. There are many pressing research challenges that need to be addressed as an important part of machine learning methods moving towards clinical practice.

6.1 Uncertainty and Fairness in Healthcare

Machine learning and probabilistic methods are widely used in different fields. Medical data is more prone to uncertainty because of the presence of noise in the data. An essential but less explored aspect is how uncertainty in data and models can be captured and analyzed. This is crucial in high-risk applications, especially in the clinical setting. For example, a doctor may perform a manual review of some highly uncertain cases based on the uncertainty of the case. Nowadays, application of novel deep learning techniques to deal such uncertainties has significantly increased [4]. However, research on the intersection of fairness and uncertainty has lacked extensive research. It is shown that uncertainty could help to identify the fairness problem in machine learning for healthcare [64, 65]. More efforts are needed to investigate the fairness problem of uncertainty in machine learning models.

6.2 Long-term Fairness in Healthcare

One more distributive justice not mentioned in Sec. 4.1 called equal outcome or equal benefit. It refers to the assurance that protected groups have equal benefit from the deployment of machine-learning models. The gap between equal allocation and equal benefit occurs when a fair decision cannot guarantee fair benefit to patients in the future. Most current research has focused on fairness issues in machine learning in static classification scenarios and has not examined how these decisions will affect the future [44]. It is often assumed that unfairness can be improved better after imposing fairness constraints on machine learning models; however, this is not the case in practice healthcare settings. Even in a one-step feedback model, ordinary standards of fairness generally do not promote improvement over time and may cause harm [61]. The key difficulty in alleviating the long-term fairness problem is to simulate the long-term dynamics and predict the future benefit [25].

Another research challenge is that the healthcare system is not a isolated system. When machine learning algorithms are embedded in clinical systems, the diagnostic decisions they make are collected and composed into new clinical data. These data in turn further influence the performance of future machine learning algorithms. This is also called a feedback loop. When bias appears in the feedback loop, it can exacerbate the bias or create new biases and further compromise the benefit of certain demographic groups. Similar feedback loop has been discussed in the context of recommend system [96]. For now, there is no research work that focuses on the long-term fairness problem in the context of the healthcare domain. We encourage more work on the long-term fairness of machine learning algorithms in healthcare, in particular on equal benefit and feedback loop fairness in clinical applications.

6.3 Fairness of Multi-modality Model for Healthcare

A research question is described as multimodal when it includes multiple data types. The human experience of the world is multi-modal. Multi-modal machine learning aims to build models that can process and correlate information from multiple modalities, thus enabling advances in artificial intelligence in understanding the world around us. One of the key driving forces of intelligent medical system is multimodal method. The combination of different modalities of healthcare data, each providing information about a patient’s treatment from a specific perspective, overlays and complements each other to further improve the accuracy of diagnosis and treatment. For example, the visual quest answering (VQA) task [6] combines computer vision (CV) and natural language processing (NLP), and the model can answer relevant questions based on medical images and clinical notes [60]. However, multi-modal models face more serious bias and fairness issues than uni-modal models, despite improvements in performance [12]. Only a few works have focused on fairness issues in multi-modality in healthcare systems [17]. The forms in which bias exists vary across modality data, as do the methods used to mitigate it. As previous work has focused on the fairness problem in uni-modal data,



we encourage the discovery and mitigation of bias in healthcare in multimodal machine learning.

6.4 Ethical Machine Learning

The concepts and attributes of the ethical scope of machine learning encompass fairness, interpretability, and privacy protection. Of these, the study of model interpretability has evolved into a separate research area [32]. It is dedicated to unlocking the black-box nature of models, especially deep learning models, and thus helping to increase trust in model decisions. This is particularly important for applications in high-risk scenarios, such as healthcare. Interpretable techniques can also be uniquely advantageous when combined with fair machine learning in healthcare scenarios. Interpretability techniques can be used to audit the failure modes of models and reveal possible biases in the models. For example, correlations site-specific staining variability and ethnicity was found in TCGA through attention-based interpretability method [48]. At present, the intersection of interpretability, fairness and machine learning in health research remains fairly narrow, with most of the current research on trustworthy AI has been conducted from a single dimension, and only a few works have examined cross-sections of how feature importance and interpretability transfer across protected subgroups. These dimensions are both consistent to a certain extent and constrained by each other. Thus, integrated mechanisms, algorithms, models and frameworks for ethical machine learning that integrate fairness, interpretability and privacy protection are worth investigating.

7 Conclusions

In this survey report, we provide an overview of current advances in machine learning for fairness in healthcare. Specifically, we first summarize the potential sources of bias in healthcare according to the different stages of machine learning development. We then sort out the fairness metrics applicable to healthcare under the concept of distributive justice. In addition, we outline what has been done to mitigate fairness issues in healthcare at different stages of machine learning development and provide a more fine-grained analysis of these efforts. Finally, we suggest some possible future work and directions worth exploring to address the issue of fairness in machine learning in healthcare that we have not discussed in the previous sections.

References

- [1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.
- [2] A. S. Adamson and A. Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.
- [3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach. A reductions approach to fair classification. *ArXiv*, abs/1803.02453, 2018.
- [4] R. Alizadehsani, M. Roshanzamir, S. Hussain, A. Khosravi, A. Koohestani, M. H. Zangoeei, M. Abdar, A. Beykikhoshk, A. Shoeibi, A. Zare, M. Panahiazar, S. Nahavandi, D. Srinivasan, A. F. Atiya, and U. R. Acharya. Handling of uncertainty in medical data using machine learning and probability theory techniques: a review of 30 years (1991–2020). *Annals of Operations Research*, pages 1 – 42, 2021.
- [5] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. Publicly available clinical bert embeddings. *ArXiv*, abs/1904.03323, 2019.
- [6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [7] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences*, 11:5088, 2021.

- [8] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion*, 58:82–115, 2020.
- [9] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. J. S. Litjens, J. A. van der Laak, M. Hermesen, Q. F. Manson, M. C. A. Balkenhol, O. G. F. Geessink, N. Stathonikos, M. C. van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H. Lin, P.-A. Heng, C. Hass, E. Bruni, Q. K.-S. Wong, U. Halici, M. Ü. Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Z. Kraus, M. Shaban, N. M. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvoori, K. Liimatainen, S. Albarqouni, B. Mungal, A. A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. A. Phoulady, V. A. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. del Milagro Fernández-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318:2199–2210, 2017.
- [10] R. A. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50:3 – 44, 2018.
- [11] K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23, 2021.
- [12] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D’Mello. Bias and fairness in multimodal machine learning: A case study of automated video interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.
- [13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.
- [14] G. Campanella, M. G. Hanna, L. Geneslaw, A. P. Mirafior, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, pages 1–9, 2019.
- [15] S. Caton and C. Haas. Fairness in machine learning: A survey. *ArXiv*, abs/2010.04053, 2020.
- [16] N. Chawla, K. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002.
- [17] J. Chen, I. Berlot-Attwell, S. Hossain, X. Wang, and F. Rudzicz. Exploring text specific and blackbox fairness algorithms in multimodal clinical nlp. *ArXiv*, abs/2011.09625, 2020.
- [18] R. J. Chen, T. Y. Chen, J. Lipková, J. J. Wang, D. F. K. Williamson, M. Y. Lu, S. Sahai, and F. Mahmood. Algorithm fairness in ai for medicine and healthcare. *ArXiv*, abs/2110.00603, 2021.
- [19] P. Cheng, W. Hao, S. Yuan, S. Si, and L. Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders, 2021.
- [20] Y. Choi, C. Y.-I. Chiu, and D. Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41, 2016.
- [21] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163, 2017.
- [22] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Fair regression with wasserstein barycenters. *ArXiv*, abs/2006.07286, 2020.

- [23] R. Correa, J. J. Jeong, B. Patel, H. Trivedi, J. W. Gichoya, and I. Banerjee. Two-step adversarial debiasing with partial learning - medical image case-studies. *ArXiv*, abs/2111.08711, 2021.
- [24] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.
- [25] 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*, pages 525–534, 2020.
- [26] A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77:605–610, 1982.
- [27] C. Denis, R. Elie, M. Hebiri, and F. Hu. Fairness guarantee in multi-class classification. 2021.
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
- [29] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth. Allotaxonomy and rank-turbulence divergence: a universal instrument for comparing complex systems. *arXiv preprint arXiv:2002.09770*, 2020.
- [30] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.
- [31] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017.
- [32] M. Du, N. Liu, and X. Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.
- [33] M. Du, N. Liu, F. Yang, and X. Hu. Learning credible deep neural networks with rationale regularization. *2019 IEEE International Conference on Data Mining (ICDM)*, pages 150–159, 2019.
- [34] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. *ArXiv*, abs/1104.3913, 2012.
- [35] G. G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nat. Mach. Intell.*, 3:620–631, 2021.
- [36] G. J. Escobar, B. J. Turk, A. I. Ragins, J. Ha, B. Hoberman, S. M. Levine, M. A. Balleca, V. X. Liu, and P. Kipnis. Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. *Journal of hospital medicine*, 11 Suppl 1:S18–S24, 2016.
- [37] D. Fan, Y. Wu, and X. Li. On the fairness of swarm learning in skin lesion classification. *ArXiv*, abs/2109.12176, 2021.
- [38] A. W. Flores, K. A. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to ”machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks”. *Federal Probation*, 80:38, 2016.
- [39] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. E. Gross, S. O. Sumer, Y. Sun, A. S. Jacobsen, R. Sinha, E. Larsson, E. G. Cerami, C. Sander, and N. D. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cbiportal. *Science Signaling*, 6:pl1 – pl1, 2013.
- [40] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness, 2019.

- [41] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.
- [42] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.
- [43] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.
- [44] H. Heidari, V. Nanda, and K. P. Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning, 2019.
- [45] K. C. Heslin, P. L. Owens, Z. Karaca, M. L. Barrett, B. J. Moore, and A. Elixhauser. Trends in opioid-related inpatient stays shifted after the us transitioned to icd-10-cm diagnosis coding in 2015. *Medical Care*, 55:918–923, 2017.
- [46] G. E. Hinton. Deep learning-a technology with the potential to transform health care. *JAMA*, 320:11101–1102, 2018.
- [47] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, 9, 2019.
- [48] F. M. Howard, J. M. Dolezal, S. E. Kochanny, J. J. Schulte, H. I.-H. Chen, L. R. Heij, D. Huo, R. Nanda, O. I. Olopade, J. N. Kather, N. A. Cipriani, R. L. Grossman, and A. T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12, 2021.
- [49] J. A. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. Langlotz, B. N. Patel, M. P. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.
- [50] S. Jabbour, D. F. Fouhey, E. A. Kazerooni, M. W. Sjoding, and J. Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *MLHC*, 2020.
- [51] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.
- [52] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. ying Deng, R. G. Mark, and S. Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *ArXiv*, abs/1901.07042, 2019.
- [53] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33, 2011.
- [54] S. Khadanga, K. Aggarwal, S. R. Joty, and J. Srivastava. Using clinical notes with time series data for icu management. *ArXiv*, abs/1909.09702, 2019.
- [55] M. Kuppler, C. Kern, R. L. Bach, and F. Kreuter. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *arXiv preprint arXiv:2105.01441*, 2021.
- [56] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NIPS*, 2017.
- [57] G. H. Kwak and P. Hui. Deephealth: Review and challenges of artificial intelligence in health informatics. *arXiv: Learning*, 2019.
- [58] M. W. Y. Lam, C.-Y. Chen, Z. Li, A. R. Martin, J. Bryois, X. Ma, H. A. Gaspar, M. Ikeda, B. Benyamin, B. C. Brown, R. Liu, W. Zhou, L. li Guan, Y. Kamatani, S.-W. Kim, M. Kubo, A. A. Kusumawardhani, C.-M. Liu, H. Ma, S. Periyasamy, A. Takahashi, Z. Xu, H. Yu, F. Zhu, W. J. Chen, S. V. Faraone, S. J. Glatt, L. He, S. E. Hyman, H.-G. Hwu, S. Mccarroll, B. M. Neale, P. Sklar, D. B. Wildenauer, X. Yu,

- D. Zhang, B. J. Mowry, J. Lee, P. A. Holmans, S. Xu, P. F. Sullivan, S. Ripke, M. C. O'Donovan, M. J. Daly, S. Qin, P. C. Sham, N. Iwata, K. S. Hong, S. G. Schwab, W. Yue, M. T. Tsuang, J. Liu, X. Ma, R. S. Kahn, Y. Shi, and H. Huang. Comparative genetic architectures of schizophrenia in east asian and european populations. *Nature genetics*, 51:1670 – 1678, 2019.
- [59] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America*, 117:12592 – 12594, 2020.
- [60] Z. Lin, D. Zhang, Q. Tac, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge. Medical visual question answering: A survey. *arXiv preprint arXiv:2111.10056*, 2021.
- [61] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. *ArXiv*, abs/1803.04383, 2018.
- [62] Q. Liu, L. Yu, L. Luo, Q. Dou, and P.-A. Heng. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, 39:3429–3440, 2020.
- [63] C. Log'e, E. L. Ross, D. Y. A. Dadey, S. Jain, A. Saporta, A. Y. Ng, and P. Rajpurkar. Q-pain: A question answering dataset to measure social bias in pain management. *ArXiv*, abs/2108.01764, 2021.
- [64] C. Lu, A. Lemay, K. Chang, K. Hoebel, and J. Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. *ArXiv*, abs/2109.04392, 2021.
- [65] C. Lu, A. Lemay, K. Hoebel, and J. Kalpathy-Cramer. Evaluating subgroup disparity using epistemic uncertainty in mammography. *ArXiv*, abs/2107.02716, 2021.
- [66] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.
- [67] N. Mehrabi, F. Morstatter, N. A. Saxena, K. Lerman, and A. G. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54:1 – 35, 2021.
- [68] C. Meng, L. Trinh, N. Xu, and Y. Liu. Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *ArXiv*, abs/2102.06761, 2021.
- [69] J. R. Minot, N. Cheney, M. Maier, D. C. Elbers, C. M. Danforth, and P. S. Dodds. Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance, 2021.
- [70] J. R. Minot, N. Cheney, M. E. Maier, D. C. Elbers, C. M. Danforth, and P. S. Dodds. Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance. *ArXiv*, abs/2103.05841, 2021.
- [71] M. Nguyen. Predicting cardiovascular risk using electronic health records. 2019.
- [72] R. B. Parikh, S. Teeple, and A. S. Navathe. Addressing bias in artificial intelligence in health care. *JAMA*, 2019.
- [73] A. Paviglianiti and E. Pasero. Vital-ecg: a de-bias algorithm embedded in a gender-immune device. *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, pages 314–318, 2020.
- [74] S. R. Pfohl, T. Duan, D. Y. Ding, and N. H. Shah. Counterfactual reasoning for fair clinical risk prediction. *ArXiv*, abs/1907.06260, 2019.
- [75] S. R. Pfohl, B. J. Marafino, A. Coulet, F. Rodriguez, L. P. Palaniappan, and N. H. Shah. Creating fair models of atherosclerotic cardiovascular disease risk. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

- [76] G. Pleiss, M. Raghavan, F. Wu, J. M. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *NIPS*, 2017.
- [77] E. Puyol-Antón, B. Ruijsink, J. M. Harana, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, P. J. Chowieńczyk, and A. P. King. Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. In *medRxiv*, 2021.
- [78] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.
- [79] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169:866–872, 2018.
- [80] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.
- [81] J.-F. Rajotte, S. Mukherjee, C. Robinson, A. Ortiz, C. West, J. L. Ferres, and R. T. Ng. Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary. *arXiv preprint arXiv:2101.07235*, 2021.
- [82] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [83] B. Scherrer, A. Gholipour, and S. Warfield. Super-resolution reconstruction to increase the spatial resolution of diffusion weighted images from orthogonal anisotropic acquisitions. *Medical image analysis*, 16 7:1465–76, 2012.
- [84] L. Seyyed-Kalantari, G. Liu, M. B. A. McDermott, and M. Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 26:232–243, 2021.
- [85] P. Tiwald, A. Ebert, and D. Soukup. Representative & fair synthetic data. *ArXiv*, abs/2104.03007, 2021.
- [86] Y.-H. H. Tsai, M. Q. Ma, H. Zhao, K. Zhang, L.-P. Morency, and R. Salakhutdinov. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *arXiv preprint arXiv:2106.02866*, 2021.
- [87] C. Wachinger and M. Reuter. Domain adaptation for alzheimer’s disease diagnostics. *NeuroImage*, 139:470–479, 2016.
- [88] F. Wang and A. M. Preininger. Ai in health: State of the art, challenges, and future directions. *Yearbook of Medical Informatics*, 28:16 – 26, 2019.
- [89] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.
- [90] 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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.
- [91] J. R. Williams and N. Razavian. Towards quantification of bias in machine learning for healthcare: A case study of renal failure prediction. *ArXiv*, abs/1911.07679, 2019.

- [92] J. Wu, D. Peck, S. S. Hsieh, V. Dialani, C. D. Lehman, B. Zhou, V. Syrgkanis, L. W. Mackey, and G. Patterson. Expert identification of visual primitives used by cnns during mammogram classification. In *Medical Imaging*, 2018.
- [93] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I.-C. Chang. Deep learning of feature representation with multiple instance learning for medical image analysis. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1626–1630, 2014.
- [94] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1280–1283, 2019.
- [95] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proceedings of the 26th International Conference on World Wide Web*, 2017.
- [96] D. Zhang and J. Wang. Recommendation fairness: From static to dynamic. *arXiv preprint arXiv:2109.03150*, 2021.
- [97] H. Zhang, A. X. Lu, M. Abdalla, M. B. A. McDermott, and M. Ghassemi. Hurtful words: quantifying biases in clinical contextual word embeddings. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020.
- [98] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.
- [99] J. Xu, Y. Xiao, W. H. Wang, Y. Ning, E. A. Shenkman, J. Bian, and F. Wang. Algorithmic fairness in computational medicine. *medRxiv*, 2022.
- [100] Y. Zhou, S.-C. Huang, J. A. Fries, A. Youssef, T. J. Amrhein, M. Chang, I. Banerjee, D. Rubin, L. Xing, N. Shah, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021.