

Research Article

Fair Diagnosis: Leveraging Causal Modeling to Mitigate Medical Bias

Bowei Tian¹, Yexiao He¹, Meng Liu¹, Yucong Dai², Ziyao Wang¹, Shwai He¹, Guoheng Sun¹, Zheyu Shen¹, Wanghao Ye¹, Yongkai Wu², Ang Li¹

1. The University of Maryland, United States; 2. Clemson University, United States

In medical image analysis, model predictions can be affected by sensitive attributes, such as race and gender, leading to fairness concerns and potential biases in diagnostic outcomes. To mitigate this, we present a causal modeling framework, which aims to reduce the impact of sensitive attributes on diagnostic predictions. Our approach introduces a novel fairness criterion, Diagnosis Fairness, and a unique fairness metric, leveraging path-specific fairness to control the influence of demographic attributes, ensuring that predictions are primarily informed by clinically relevant features rather than sensitive attributes. By incorporating adversarial perturbation masks, our framework directs the model to focus on critical image regions, suppressing bias-inducing information. Experimental results across multiple datasets demonstrate that our framework effectively reduces bias directly associated with sensitive attributes while preserving diagnostic accuracy. Our findings suggest that causal modeling can enhance both fairness and interpretability in AI-powered clinical decision support systems.

1. Introduction

Medical image analysis driven by deep learning has achieved impressive success, often reaching or exceeding human expert-level diagnostic performance across various tasks^{[1][2][3]}. While these advanced medical image models have significantly improved the accuracy and efficiency of medical diagnosis, they have also raised concerns about the fairness and reliability of AI-driven decisions. It has been revealed that many medical AI systems unintentionally incorporate sensitive demographic attributes (e.g., race, gender, age) into their decision-making process, potentially leading to biased predictions and compromised healthcare equality^{[4][5]}. This systematic bias undermines the

trustworthiness of medical AI systems and raises ethical concerns regarding their deployment in real-world clinical settings.

The core of this issue lies in the challenge of controlling the influence of sensitive attributes within medical images. For example, in chest X-ray diagnosis, models may rely on anatomical variations associated with gender or race instead of focusing on pathology-related features^{[6][7][8]}. Although one might consider omitting sensitive information, models often infer these demographics from correlated image patterns, perpetuating the bias^[9]. Moreover, removing sensitive features could compromise diagnostic precision, as these features often correlate with critical diagnostic signals^{[10][11][12]}. Therefore, achieving the optimized balance between fairness and accuracy is crucial for ethical and effective AI in medical imaging.

Previous studies on mitigating bias in medical AI applications can be broadly categorized into data-centric and algorithmic approaches. Data-centric approaches focus on improving the diversity and balance of datasets to reduce bias. For instance, Burlina et al. employed generative methods for data augmentation, minimizing diagnostic disparities in retinal image classification between light-skinned and dark-skinned individuals^[13]. Other approaches, such as resampling strategies, have been employed to address dataset imbalance issues, ensuring fairer representation of underrepresented subgroups^{[14][15]}.

Algorithmic methods address fairness during the model training process by modifying the training objective or employing specialized architectures. For example, Paul et al. proposed the Training and Representation Alteration (TARA) framework^[14], which adapts domain generalization techniques to improve fairness across demographic groups. Similarly, Zhou et al. developed multimodal methods for pulmonary embolism detection^[16], demonstrating the impact of architectural choices on fairness outcomes. Another important debiasing technique is adversarial learning, such as adversarial debiasing (AD)^[10] and fairness-aware adversarial perturbation (FAAP)^[17]. AD^[10] uses adversarial learning to mitigate biases within the model by designing the training process as an adversarial game, where a secondary network attempts to detect protected attributes (e.g., gender, race) from the primary model's output. FAAP^[17] takes a distinct approach by generating adversarial perturbations that directly incorporate fairness constraints; these perturbations challenge the model's robustness while assessing and mitigating its bias.

While these methods succeed in domains like face recognition and tabular classification, they often face limitations in medical imaging, where sensitive and diagnostic attributes are typically entangled, making it difficult to separate demographic influence from clinically relevant information. Most existing approaches attempt to either blind models to sensitive attributes or process datasets without fully understanding the causal relationships between sensitive attributes, image features, and diagnostic outcomes. This limitation can result in an inability to distinguish genuine clinical differences from undesirable demographic biases, ultimately compromising diagnostic accuracy. Therefore, achieving fairness in medical AI requires rethinking the existing bias-mitigation strategies and evaluations.

To address these challenges, we propose a novel algorithmic method utilizing causal modeling that explicitly distinguishes the direct effect and indirect effect caused by the sensitive attribute. Theoretically, our approach leverages structural causal models to identify and isolate the direct influence of sensitive attributes on predictions. Practically, we deploy adversarial training and rethink the rationale behind the algorithm, creating unique causal-based fairness concepts and metrics. To the best of our knowledge, previous fairness metrics are agnostic to causal modeling. Our primary contributions can be summarized as follows:

- We propose Diagnosis Fairness, a novel fairness criterion rooted in causal modeling, designed to ensure that diagnostic decisions are driven primarily by medically relevant information. We further introduce the Approximate Diagnosis Fairness metric to evaluate the criterion effectively.
- We translate this criterion into practice by leveraging conditional mutual information and incorporating adversarial training and data utility enhancement to achieve improved fairness with minimal impact on diagnostic accuracy.
- We conduct extensive experiments, including baseline comparisons, data utility evaluations, explainability analyses, and ablation studies across multiple real-world medical datasets and downstream tasks. The results demonstrate the effectiveness and robustness of our approach.

2. Preliminaries

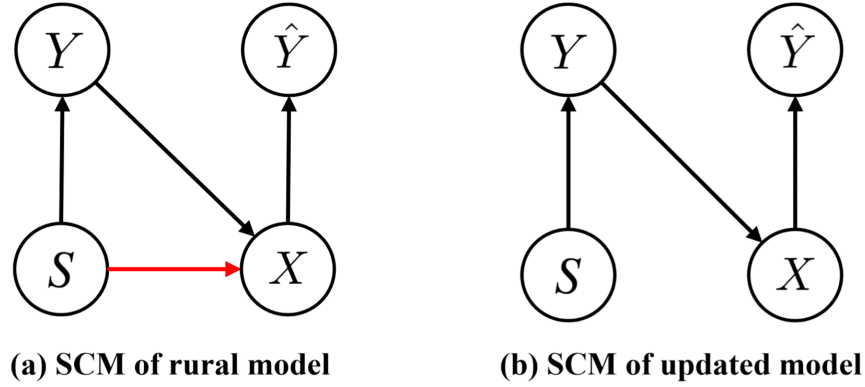


Figure 1. The SCM of our framework, where arrows represent causal paths. Here, S denotes the sensitive attribute, Y represents the true diagnosis, $\hat{Y}$ is the predicted diagnosis, and X is the input modality (e.g., X-ray images). Red paths indicate biased paths that will be removed to improve diagnosis fairness.

2.1. Causal Effect

The concept of causal effect^{[18][19][20]} refers to the causal influence that one variable (the cause variable) exerts on another (the outcome variable). Unlike mere correlation, a causal effect signifies that changes in the cause variable lead to changes in the outcome variable through a direct causal mechanism. We now define the specific causal effects within the Structural Causal Model (SCM)^[21] for our framework, as illustrated in Fig. 1.

Definition 1 (Total Causal Effect^[18]). The Total Causal Effect (TCE) of a change in the value of the sensitive attribute S from S^- to S^+ on $\hat{Y}$ is given by:

$$\text{TCE}(S) = P(\hat{Y}|S^+) - P(\hat{Y}|S^-). \quad (1)$$

The TCE measures the influence of S on $\hat{Y}$ as the effect propagates along all causal paths from S to $\hat{Y}$. However, if we consider the influence along only a subset of causal paths from S to $\hat{Y}$, we refer to the resulting effect as the path-specific effect, defined below.

Definition 2 (Direct Effect^{[19][20]}). Given the direct path $\pi_d = \{S \rightarrow X \rightarrow \hat{Y}\}$ and the indirect path $\pi_i = \{S \rightarrow Y \rightarrow X \rightarrow \hat{Y}\}$, the Direct Effect (DE) represents the path-specific effect of S^+ along π_d ,

with S^- along π_i :

$$DE(S) = P(\hat{Y}|S_{\pi_d}^+, S_{\pi_i}^-) - P(\hat{Y}|S^-), \quad (2)$$

where $P(\hat{Y}|S_{\pi_d}^+, S_{\pi_i}^-)$ represents the post-intervention distribution of $\hat{Y}$ with the intervention $do(S^+)$ affecting only the direct path π_d , while the reference intervention $do(S^-)$ influences the indirect path π_i .

Definition 3 (Indirect Effect [19][20]). Utilizing the concept of TCE and DE within the causal graph, we define the Indirect Effect (IE) as:

$$\begin{aligned} IE(S) &= TCE(S) - DE(S) \\ &= (P(\hat{Y}|S^+) - P(\hat{Y}|S^-)) \\ &\quad - (P(\hat{Y}|S_{\pi_d}^+, S_{\pi_i}^-) - P(\hat{Y}|S^-)) \\ &= P(\hat{Y}|S^+) - P(\hat{Y}|S_{\pi_d}^+, S_{\pi_i}^-), \end{aligned} \quad (3)$$

where IE captures the effect from S that propagated through the indirect path π_i .

2.2. Fairness Concept

Traditional fairness ensures that predictive models treat individuals equally across different sensitive groups, such as race, gender, and age^{[22][23]}. From a causal perspective, fairness can be depicted as $TCE(S) = 0$ defined in Eq. 1. The criterion showcases that no explicit bias caused or conditioned by S should be allowed. However, the proposed concept neglects that true diagnosis Y can be impacted by S , which is an indirect effect that contributes to the predicted diagnosis $\hat{Y}$. For example, doctors can be more confident diagnosing a specific gender if the disease only happens to that gender. In this case, the effect of S is devoid of bias and should be considered a contribution. However, resorting directly from S to decide is obviously biased because S should not be regarded as the direct cause of a disease. Considering both the contribution and the bias exhibited in S , we formally define the **Diagnosis Fairness (DF)**: $DE(S) = 0$. This criterion proposes to prune π_d to exclude bias while allowing for S to contribute to the diagnosis within π_i .

3. Method

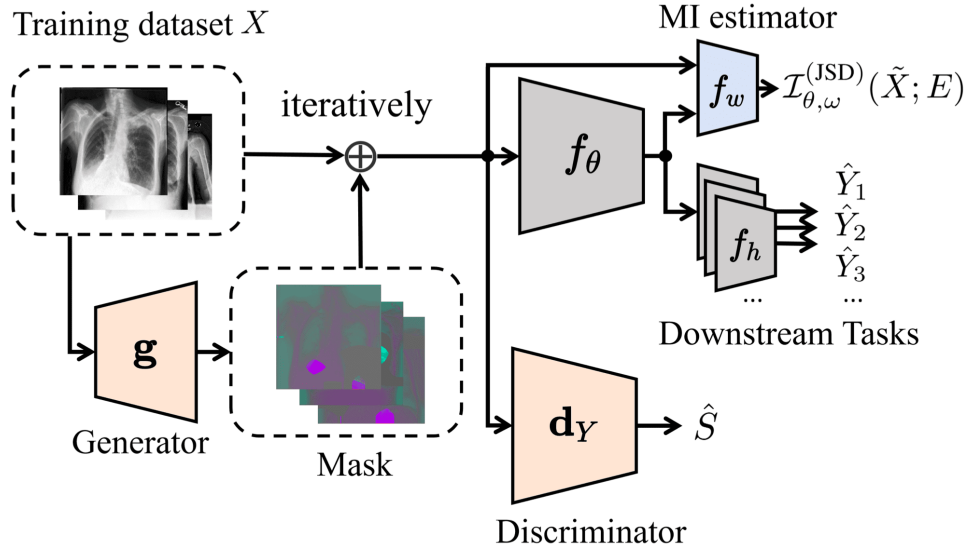


Figure 2. Overview of the Fair Diagnosis framework. The deployed models, f_θ and f_h (in grey), are fixed after pretraining. Data utility enhancement (in blue) is applied during pretraining, while adversarial training (in orange) ensures fair masking and maintains original model performance.

Figure 2 provides an overview of the Fair Diagnosis framework. Our approach uses adversarial masks to suppress the direct effect of sensitive attributes on the model’s predictions. The framework integrates adversarial training and mask generation, supported by data utility enhancement during the pretraining phase. This method effectively minimizes conditional mutual information, $\mathcal{I}(S; X|Y) = 0$, thereby enhancing Diagnosis Fairness (DF). To clarify the theoretical motivation behind this approach, we provide a proof in Section 3.3 demonstrating that achieving $\mathcal{I}(S; X|Y) = 0$ is sufficient to fulfill the DF criterion.

3.1. Pretraining

Mutual Information (MI) [24] measures the amount of information shared between two random variables. Inspired by this, we propose a data utility pretraining phase based on MI to enhance the generalization of our model for downstream tasks. This phase aims to guide the embedding to retain

essential information from the source domain, allowing the learned representations to generalize effectively across multiple tasks. Specifically, we define the embedding as:

$$E = f_{\theta}(\tilde{X}), \quad (4)$$

where $\tilde{X} = X \oplus \mathbf{g}(X)$ belongs to the source domain, and the operation $\oplus$ means iteratively adding with both upper and lower constraints. After that, we employ the Jensen-Shannon Mutual Information (Jensen-Shannon MI) estimator [25][26][27] to maximize the mutual information between the representations learned from $\tilde{X}$ and E . By using Jensen-Shannon MI, we can approximate a lower bound on the mutual information, which ensures that the most relevant features from the source domain are retained during the pretraining process.

Jensen-Shannon Mutual Information (MI), acting as the lower bound of MI, is defined as follows:

$$\begin{aligned} \mathcal{I}(\tilde{X}; E) &\geq \mathcal{I}_{\theta, \omega}^{(\text{JSD})}(\tilde{X}; E) \\ &:= \mathbb{E}_{\tilde{X}, E} \left[-\sigma \left(-f_{\omega}(\tilde{X}, E) \right) \right] - \mathbb{E}_{\tilde{X}, E'} \left[\sigma \left(f_{\omega}(\tilde{X}, E') \right) \right] \end{aligned} \quad (5)$$

where $\mathcal{I}(\cdot)$ is MI, f_{ω} is a neural network estimator parameterized by ω , and $\sigma(z) = \log(1 + e^z)$ is the softplus function. Here, $E = f_{\theta}(\tilde{X})$ is the positive embedding from the target domain, and E' is a shuffled version of E , acting as the negative embedding. Hence, to maximally retain the original information, the feature extractor and the mutual information estimator can be optimized using Eq. 6:

$$\arg \max_{\theta} \max_{\omega} \mathcal{I}_{\theta, \omega}^{(\text{JSD})}(\tilde{X}; E). \quad (6)$$

By maximizing this objective, we ensure that the learned representations E effectively transfer meaningful information from $\tilde{X}$ to the target domain, promoting robust data utility. Therefore, we can conveniently extend to multiple tasks by applying individual fine-tuning on $\hat{Y}_i = f_h(E)_i$.

3.2. Adversarial Training

Next, let us explore how DF is approximated via adversarial training between the generator $\mathbf{g}$ and the conditioned discriminator $\mathbf{d}_Y$. Specifically, the generator $\mathbf{g}$ is used to generate the masks. After adding the masks to the original training set X , the modified data $\tilde{X}$ is passed to the discriminator $\mathbf{d}_Y$. The discriminator then predicts $\hat{S}$ while conditioned on Y , where the conditioning means that we simultaneously input Y to $\mathbf{d}_Y$ to make it directly learn the patterns in Y . In other words, the input and output of $\mathbf{d}_Y$ can be described as:

$$\hat{S} = \mathbf{d}_Y(\tilde{X}, Y). \quad (7)$$

Acting as $\mathbf{d}_Y$'s opponent, the generator $\mathbf{g}$ will generate masks that fail $\mathbf{d}_Y$ by ensuring that S and X are as independent as possible under the condition of Y ; otherwise, $\mathbf{d}_Y$ will be able to predict S with ease. Theoretically, the objective of $\mathbf{g}$ is:

$$\mathcal{I}(S; X | Y) = 0. \quad (8)$$

To achieve this objective, we consequently implement concrete optimization techniques. The optimization functions in this adversarial setting guide the interaction between $\mathbf{g}$ and $\mathbf{d}_Y$. Specifically:

Generator Loss: The generator therefore aims to create perturbations that hinder the discriminator from correctly predicting $\hat{S}$, which can be achieved by minimizing the mutual information $\mathcal{I}(\hat{S}; S)$. However, merely minimizing it will push the latent representations towards the opposite side of the sensitive attribute, e.g., female flips to male. Therefore, we further add a regularization that lets $\mathbf{d}_Y$ make random guesses on the perturbed images by increasing the entropy $\mathcal{H}$ of the protected attribute:

$$\mathcal{L}_{\mathbf{g}}^{fair} = \mathcal{I}(\hat{S}; S) - \alpha \mathcal{H}(\hat{S}), \quad (9)$$

where $\alpha > 0$ is a relatively small value that controls the regularization of the entropy loss. Simultaneously, we want the generator to maintain the accuracy of the deployed model. Therefore, the overall $\mathcal{L}_{\mathbf{g}}$ can be calculated as:

$$\mathcal{L}_{\mathbf{g}} = \mathcal{L}_{\mathbf{g}}^{fair} - \beta \mathcal{I}(\hat{Y}; Y), \quad (10)$$

where $\beta > 0$ balances the accuracy and fairness trade-off.

Discriminator Loss: The discriminator tries to maximize its ability to predict the protected attribute $\hat{S}$ from the perturbed images, given the prior knowledge of diagnosis Y (shown in Eq. 7), formulated as:

$$\mathcal{L}_{\mathbf{d}_Y} = -\mathcal{I}(\hat{S}; S), \quad (11)$$

where the objective of $\mathcal{L}_{\mathbf{d}_Y}$ acts as the adversarial of the first term of $\mathcal{L}_{\mathbf{g}}^{fair}$, which showcases how the adversarial training works. Consequently, the iterative adversarial training ensures that the generator learns to create fair perturbations (by fooling the discriminator) while preserving the original model performance, as the discriminator tries to distinguish which sensitive group each sample belongs to. Therefore, the objectives of our method can be formulated as follows:

$$\arg \max_{\mathbf{g}} \min_{\mathbf{d}_Y} -\mathcal{I}(\hat{S}; S) + \alpha \mathcal{H}(\hat{S}) + \beta \mathcal{I}(\hat{Y}; Y), \quad (12)$$

where $\mathbf{d}_Y$ and $\mathbf{g}$ are updated alternatively during the optimization. Note that α and β are set to 0 during the updating of $\mathbf{d}_Y$ to allow $\mathbf{d}_Y$ to focus on distinguishing protected attributes. By leveraging adversarial training between $\mathbf{g}$ and $\mathbf{d}_Y$, we can reach a better DF without changing the parameters of the deployed model.

3.3. Theoretical Analysis

In this section, we analyze the theoretical relationship between the objective of the generator $\mathbf{g}$ and diagnosis fairness, as well as the feasibility of evaluating diagnosis fairness through $\mathcal{I}(S; \hat{Y} | Y) = 0$.

Theorem 1. For random variables Y, S, X and $\hat{Y}$, the conditional mutual information $\mathcal{I}(S; X | Y) = 0$ is a sufficient and not necessary condition of $DE(S) = 0$.

Proof. (Sufficiency): According to Eq. 2:

$$\begin{aligned} DE(S) &= P(\hat{Y} | S_{\pi_d}^+, S_{\pi_i}^-) - P(\hat{Y} | S_{\pi_d}^-, S_{\pi_i}^-) \\ &= \sum_Y P(\hat{Y} | X) P(X | Y, S_{\pi_d}^+) P(Y | S_{\pi_i}^-) \\ &\quad - \sum_Y P(\hat{Y} | X) P(X | Y, S_{\pi_d}^-) P(Y | S_{\pi_i}^-) \\ &= P(\hat{Y} | X) \left[\sum_Y P(X | Y, S_{\pi_d}^+) P(Y | S_{\pi_i}^-) \right. \\ &\quad \left. - \sum_Y P(X | Y, S_{\pi_d}^-) P(Y | S_{\pi_i}^-) \right], \end{aligned} \quad (13)$$

apply Bayesian Rule^[28] to the formula:

$$\begin{aligned} DE(S) &= \left[\sum_Y \frac{P(S_{\pi_d}^+ | X, Y) P(X | Y) P(Y | S_{\pi_i}^-)}{P(S_{\pi_d}^+ | Y)} \right. \\ &\quad \left. - \sum_Y \frac{P(S_{\pi_d}^- | X, Y) P(X | Y) P(Y | S_{\pi_i}^-)}{P(S_{\pi_d}^- | Y)} \right], \\ &= \left[\sum_Y \frac{P(S_{\pi_d}^+ | X, Y) P(Y | S_{\pi_i}^-)}{P(S_{\pi_d}^+ | Y)} \right. \\ &\quad \left. - \sum_Y \frac{P(S_{\pi_d}^- | X, Y) P(Y | S_{\pi_i}^-)}{P(S_{\pi_d}^- | Y)} \right]. \end{aligned} \quad (14)$$

According to the definition of $\mathcal{I}(S; X | Y) = 0$, it satisfies $(S \perp X | Y)$, which can be described as:

$$P(S | X, Y) = P(S | Y), \quad (15)$$

Let S impact on π_d , we can deduce:

$$\frac{P(S_{\pi_d}^+ | X, Y)}{P(S_{\pi_d}^+ | Y)} = \frac{P(S_{\pi_d}^- | X, Y)}{P(S_{\pi_d}^- | Y)} = C, \quad (16)$$

where C is a constant. Substitute into Eq. 14:

$$\text{DE}(S) = C \cdot \sum_Y (\text{P}(Y | S_{\pi_i}^-) - \text{P}(Y | S_{\pi_i}^-)) = 0 \quad (17)$$

(Necessity): Since Eq. 16 is a special case of Eq. 15, therefore,

$$\text{DE}(S) = 0 \implies \mathcal{I}(S; X | Y) = 0, \quad (18)$$

so $\mathcal{I}(S; X | Y) = 0$ is not a necessary condition of $\text{DE}(S) = 0$. $\square$

Theorem 1 concludes that if we meet the requirement of $\mathcal{I}(S; X|Y) = 0$, DF will be satisfied, where π_d will be pruned, showcasing that the adversarial training between $\mathbf{g}$ and $\mathbf{d}_Y$ is theoretically reaching DF. Subsequently, we further prove that $\mathcal{I}(S; \hat{Y}|Y) = 0$ can be the evaluation metric for the objective of adversarial training, shedding light on the DF evaluation.

Theorem 2. For random variables Y , S , X , and $\hat{Y}$, $\mathcal{I}(S; X|Y) = 0$ is a sufficient and not necessary condition for $\mathcal{I}(S; \hat{Y}|Y) = 0$.

Proof. (Sufficiency): Assume that X is a complete mediator between $\hat{Y}$ and (Y, S) . This assumption implies that any dependence between $\hat{Y}$ and (Y, S) can be fully mediated by X . Using the law of total probability^[29], we can express $P(\hat{Y}|Y, S)$ as follows:

$$\begin{aligned} P(\hat{Y}|Y, S) &= \int P(\hat{Y}|X, Y, S)P(X|Y, S)dX \\ &= \int P(\hat{Y}|X)P(X|Y, S)dX \\ &= \int P(\hat{Y}|X)P(X|Y)dX \\ &= P(\hat{Y}|Y). \end{aligned} \quad (19)$$

Thus, the equation holds $(S \perp \hat{Y}|Y)$, which is equivalent to $\mathcal{I}(S; \hat{Y}|Y) = 0$.

(Necessity): To show that $\mathcal{I}(S; \hat{Y}|Y) = 0$ does not necessarily imply $\mathcal{I}(S; X|Y) = 0$, consider the following:

$$P(X|Y, S) = \sum_{\hat{Y}} P(X|\hat{Y}, Y, S)P(\hat{Y}|Y, S). \quad (20)$$

While

$$\begin{aligned} P(X|Y) &= \sum_{\hat{Y}} P(X|\hat{Y}, Y)P(\hat{Y}|Y) \\ &= \sum_{\hat{Y}} P(X|\hat{Y}, Y)P(\hat{Y}|Y, S). \end{aligned} \quad (21)$$

It is generally not possible to derive Eq. 20 from Eq. 21 without assuming that $(S \perp X|Y)$. This indicates that $\mathcal{I}(S; X|Y) = 0$ is not a necessary condition for $\mathcal{I}(S; \hat{Y}|Y) = 0$. $\square$

Since directly measuring $\mathcal{I}(S; X|Y) = 0$ and DF can be inherently challenging, Theorem 2 concludes that $\mathcal{I}(S; \hat{Y}|Y) = 0$ is a sufficient condition for $\mathcal{I}(S; X|Y) = 0$, making it a practical evaluation metric. We define $\mathcal{I}(S; \hat{Y}|Y)$ as the **Approximate Diagnosis Fairness (ADF)**. Given that we aim to minimize $\mathcal{I}(S; X|Y)$ in our method and have proven that both DF and ADF are sufficient conditions for $\mathcal{I}(S; X|Y) = 0$, ADF serves as a valuable metric to measure DF in evaluations.

4. Experiments

4.1. Experiment Settings

4.1.1. Datasets

MIMIC-CXR^[30]: The MIMIC-CXR dataset, part of the MIMIC initiative, includes over 370,000 chest X-rays from more than 60,000 patients, labeled with conditions like pneumonia, pleural effusion, and lung lesions. This anonymized dataset is valuable for machine learning research, supporting tasks such as automated disease detection and clinical decision support, especially in critical care environments.

CheXpert^[31]: The CheXpert dataset, created by Stanford University, contains over 224,000 chest X-rays annotated for 14 radiological findings, including lung opacity and cardiomegaly. Known for its high-quality labels obtained through a robust NLP-based pipeline, CheXpert is widely used in AI research for disease classification and anomaly detection in medical imaging.

TCGA-LUAD^[32]: This dataset from The Cancer Genome Atlas (TCGA) project focuses on lung adenocarcinoma, providing gene expression, mutation profiles, methylation, and clinical data. It is instrumental for research in lung cancer biomarkers, diagnostics, and targeted therapies, supporting advancements in cancer genomics and precision medicine.

Dataset	Baselines	ACC%	AUC%	EO _{e-2}	DP _{e-2}	ADF _{e-3}
MIMIC	Vanilla ^[33]	85.46	63.85	5.10	6.31	10.59
	AD ^[10]	83.90	62.01	4.58	5.56	7.38
	FAAP ^[17]	83.07	61.05	4.04	5.20	6.07
	Ours	84.26	62.10	2.02	3.53	2.25
CheXpert	Vanilla ^[33]	85.87	65.93	5.05	8.86	11.18
	AD ^[10]	84.45	63.77	3.43	7.18	6.31
	FAAP ^[17]	83.50	62.55	2.86	5.40	4.84
	Ours	85.80	64.99	0.98	1.16	1.20
TCGA	Vanilla ^[33]	98.47	98.58	3.57	11.94	7.21
	AD ^[10]	97.55	97.87	3.20	9.84	4.08
	FAAP ^[17]	97.10	97.64	2.75	9.46	3.79
	Ours	97.85	98.20	1.39	6.89	1.46

Table 1. Main results on MIMIC-CXR, CheXpert, and TCGA-LUAD datasets. For simplicity, we refer to MIMIC-CXR as MIMIC, and TCGA-LUAD as TCGA in the experiments. All models are pretrained and fine-tuned on ResNet50.

4.1.2. Fairness Metrics

We evaluate fairness using both traditional metrics and our proposed ADF metric:

Demographic Parity (DP)^[23] measures the disparity in prediction rates between different sensitive groups (e.g., $S = 0$ vs. $S = 1$). Formally,

$$DP = |P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)|, \quad (22)$$

where a smaller DP value indicates fewer disparities between sensitive groups in predictions.

Equalized Opportunity (EO)^[22] is a notion of nondiscrimination with respect to a specified protected attribute, measuring the disparity in true positive rates across sensitive groups. Specifically,

$$EO = |P(\hat{Y} = 1|S = 1, Y = 1) - P(\hat{Y} = 1|S = 0, Y = 1)|. \quad (23)$$

As two of the most popular fairness metrics, DP focuses on the *positive prediction rate* across sensitive groups, while EO focuses on the *true positive rate*.

Approximate Diagnosis Fairness (ADF) quantifies the amount of mutual information between the predicted diagnosis $\hat{Y}$ and the sensitive attribute S , conditioned on the true diagnosis Y . Specifically,

$$ADF = \mathcal{I}(S; \hat{Y}|Y). \quad (24)$$

ADF captures how much uncertainty about $\hat{Y}$ can be reduced by knowing S , under the condition of Y . As proved by Theorem 1 and Theorem 2, ADF serves as an effective metric for evaluating diagnosis fairness by considering both bias and clinically relevant contributions from S , while EO and DP aim to directly exclude the effect of S from all the circumstances.

4.2. Baselines

We compare with three baseline methods: Vanilla^[33], Adversarial Debiasing (AD)^[10], and Fairness-Aware Adversarial Perturbation (FAAP)^[17], where Vanilla^[33] uses a classic training method, i.e., gradient descent to optimize; AD^[10] introduces an adversarial network that attempts to detect protected attributes (e.g., gender, race) from the primary model's output; while FAAP^[17] generates adversarial perturbations that directly incorporate fairness constraints, mitigating bias.

4.3. Main Results

The main results across the three datasets are presented in Table 1. Compared with AD, we showcase a considerable improvement in fairness, without sacrificing much accuracy; compared with FAAP, we simultaneously reach better accuracy and fairness. Therefore, these observations demonstrate the effectiveness of our method. Additionally, we observe a more distinct difference between ADF and traditional fairness metrics, including EO and DP. For example, in the TCGA-LUAD dataset, compared with Vanilla, our evaluation of ADF is lower by 5 times, while EO and DP are approximately lower by 2 times. This distinction aligns with the objective of our method and also showcases the effectiveness of the proposed fairness metric.

4.4. Data Utility Assessment

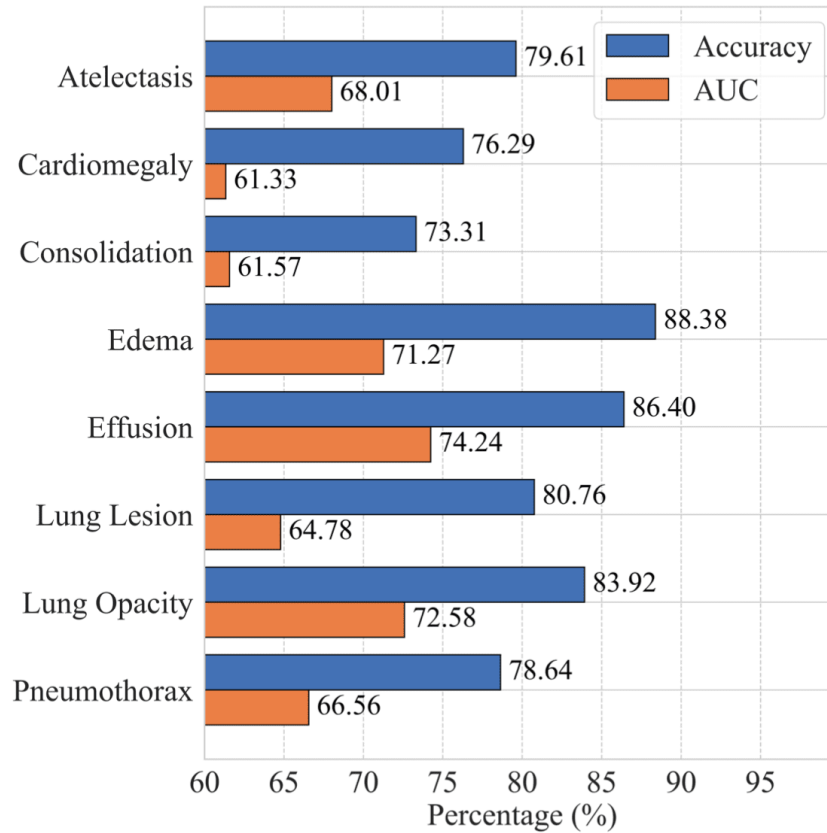


Figure 3. Data utility evaluations on the MIMIC-CXR dataset, showing the fine-tuning performance of f_h across eight different downstream tasks.

We evaluate the data utility of the MIMIC-CXR dataset after applying the MI estimator. Specifically, we start with a pretrained model for “Pneumonia” classification and subsequently fine-tune f_h across eight different downstream tasks, including the diagnosis of “Atelectasis”, “Cardiomegaly”, “Consolidation”, etc. The experimental results, presented in Fig. 3, indicate that most tasks achieve an AUC of 65% and an accuracy of 80%, demonstrating the robustness and versatility of the MI estimator across downstream tasks.

Dataset	Models	ACC _%	AUC _%	EO _{e-2}	DP _{e-2}	ADF _{e-3}
MIMIC	Resnet18	84.26	62.10	2.02	3.53	2.25
	Resnet50	84.67	62.46	1.78	2.39	1.19
	ViT	86.95	63.90	3.35	2.98	2.07
CheXpert	Resnet18	85.80	64.99	0.98	1.16	1.20
	Resnet50	86.40	66.34	1.12	1.41	1.34
	ViT	86.98	67.53	1.78	1.76	2.51
TCGA	Resnet18	97.25	98.20	1.89	7.06	1.64
	Resnet50	97.87	98.23	1.40	6.95	1.51
	ViT	98.45	99.03	2.02	8.97	1.98

Table 2. Ablation studies on different models.

4.5. Explainability Analysis

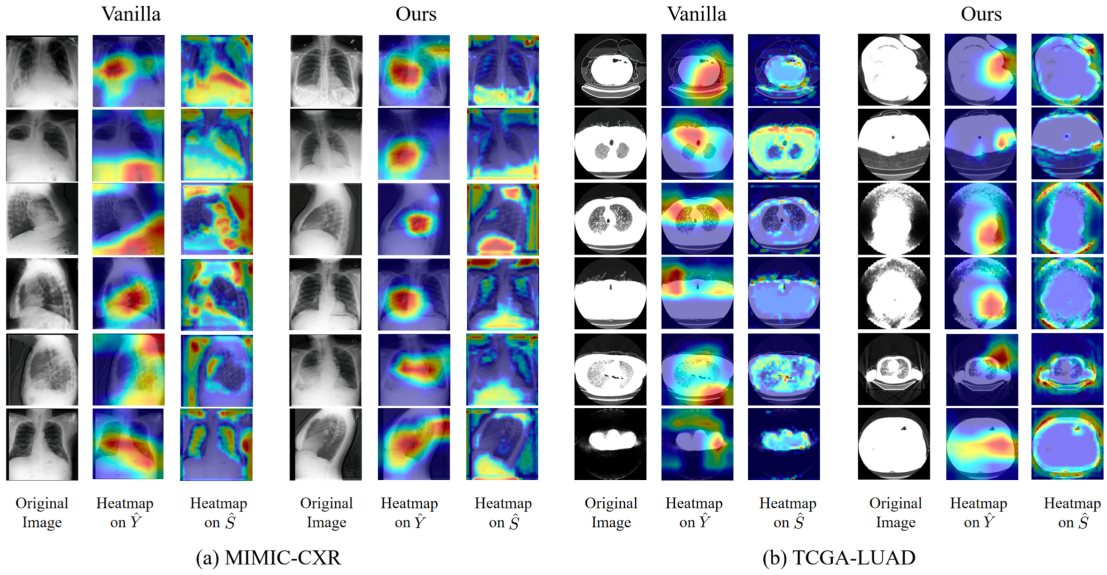


Figure 4. Explainability analysis on both MIMIC-CXR and TCGA-LUAD datasets. GradCAM^[34] is applied to the last convolution layer of f_θ when applied to $\hat{Y}$, and to $\mathbf{d}_Y$ when applied to $\hat{S}$.

We perform an explainability analysis on the MIMIC-CXR and TCGA-LUAD datasets, where GradCAM^[34] is utilized to generate heatmaps. These heatmaps visualize the regions that contribute most significantly to the model’s predictions, for both the predicted diagnosis $\hat{Y}$ and the sensitive attribute prediction $\hat{S}$. The results in Fig. 4 demonstrate the difference between Vanilla and our Fair Diagnosis framework. We observe that in Vanilla, the heatmap on $\hat{Y}$ is highly overlapped with that on $\hat{S}$, indicating the bias from sensitive attributes. For example, in the MIMIC-CXR dataset, the model may focus on the breast region in female patients, meaning that any variation in “breast shape” could inadvertently affect the diagnosis. This overlap reveals an implicit bias, as the model leverages sensitive attribute information directly in making predictions. In contrast, our method showcases a clear separation between the regions highlighted for $\hat{Y}$ and $\hat{S}$ across both datasets. This separation suggests that our method effectively disentangles the information related to the sensitive attribute $\hat{S}$ from the diagnosis prediction $\hat{Y}$, thereby pruning the direct effect of S . By minimizing the overlap, our framework ensures that predictions are driven by medically relevant features rather than sensitive attributes.

4.6. Ablation Studies

4.6.1. Model Ablation Study

We conducted a model ablation study on three different architectures: ResNet18, ResNet50, and Vision Transformers (ViTs). The results are presented in Table 2. We observe that ViTs perform slightly better in terms of accuracy and AUC but with a minor trade-off in fairness metrics. This suggests that our method can enhance fairness across various neural network architectures without significantly compromising performance. For the remaining experiments, we use ResNet50 as it provides a balanced trade-off between performance and fairness.

4.6.2. Parameter Ablation Study

We further examine the effect of different hyperparameters on model performance and fairness.

Dataset	η	ACC _%	AUC _%	EO _{e-2}	DP _{e-2}	ADF _{e-3}
MIMIC	0	85.46	63.85	5.10	6.31	10.59
	0.1	84.75	62.90	3.48	3.75	3.98
	0.2	84.26	62.10	2.02	3.53	2.25
	0.3	81.00	60.27	2.25	3.72	2.33
	0.4	80.70	59.95	2.67	4.50	2.59
CheXpert	0	85.87	65.93	5.05	8.86	11.18
	0.1	85.16	64.50	4.90	4.83	7.82
	0.2	84.59	63.11	3.47	2.64	3.24
	0.3	84.02	63.00	2.52	1.02	3.68
	0.4	82.05	61.07	4.58	3.63	9.11
TCGA	0	98.47	98.58	3.57	11.94	7.21
	0.1	98.16	98.43	2.38	9.09	5.14
	0.2	97.85	98.20	1.39	6.89	1.46
	0.3	97.16	98.09	1.69	7.69	2.82
	0.4	96.55	97.63	1.78	7.95	2.98

Table 3. Ablation studies on different noise strength η .

Ablation on Noise Strength η : Table 3 shows the effect of varying the noise strength η in the mask generation. We observe that as η increases, the performance gradually decreases. This decline suggests that larger η values result in stronger masking, which filters out more information, potentially removing some relevant diagnostic features. As for fairness, EO, DP, and ADF metrics initially improve with a slight increase in η , indicating that a small amount of noise helps reduce bias.

However, when η becomes too large, fairness begins to deteriorate. This is likely because the generator struggles to optimize effectively under strong masking, thus impacting both performance and fairness. Based on these findings, we set $\eta = 0.2$ in the remaining experiments to balance performance and fairness.

Dataset	α	ACC _%	AUC _%	EO _{$e-2$}	DP _{$e-2$}	ADF _{$e-3$}
MIMIC	0	86.53	63.05	5.56	3.58	2.32
	1	86.46	62.79	4.50	2.54	1.62
	2	86.58	63.14	3.84	1.89	1.41
	3	86.92	64.68	3.55	1.22	1.20
	4	85.99	63.02	3.51	1.02	1.13
CheXpert	0	85.41	64.93	2.42	4.48	5.12
	1	85.21	65.24	1.33	3.41	3.36
	2	85.52	65.45	1.30	3.08	3.12
	3	85.29	65.18	1.22	2.84	2.26
	4	85.60	65.52	1.00	1.67	1.52
TCGA	0	98.42	98.39	2.38	9.54	4.79
	1	98.70	98.51	1.89	7.74	3.42
	2	98.02	97.94	1.75	7.50	1.50
	3	98.55	98.25	1.43	7.01	1.47
	4	98.19	98.01	1.39	6.89	1.46

Table 4. Ablation studies on different α .

Ablation on Entropy Loss Regularization α : Table 4 demonstrates the impact of varying the entropy

loss regularization parameter α . As α increases, we observe a gradual decrease in performance, while fairness metrics improve. This trade-off occurs because the entropy loss $\mathcal{H}(\hat{S})$ encourages the generator to produce predictions that make the discriminator's task of predicting S more challenging, which promotes fairness. Based on this trade-off, we set $\alpha = 1$ in the remaining experiments to achieve an optimal balance between performance and fairness.

Dataset	β	ACC%	AUC%	EO $_{e-2}$	DP $_{e-2}$	ADF $_{e-3}$
MIMIC	0	80.69	54.65	1.08	2.63	0.98
	1	84.98	62.34	1.39	2.78	1.87
	2	85.03	62.87	2.03	3.34	2.55
	3	86.98	64.40	5.56	4.73	5.06
	4	87.79	64.80	7.58	5.31	5.50
CheXpert	0	71.15	59.75	0.54	0.42	1.31
	1	84.59	65.13	3.32	2.02	1.54
	2	85.90	65.40	3.50	3.03	2.04
	3	86.73	65.89	3.75	3.36	2.45
	4	87.74	66.65	4.50	3.48	5.25
TCGA	0	72.70	92.01	1.03	6.45	1.42
	1	97.85	98.20	1.39	6.89	1.46
	2	98.16	98.63	1.97	9.30	2.27
	3	98.30	98.71	2.03	9.35	3.55
	4	98.33	99.24	4.76	9.97	5.81

Table 5. Ablation studies on different β .

Ablation on Loss Regularization β : The results are presented in Table 5. This parameter controls the weight of the generator loss term $\mathcal{I}(\hat{Y}, Y)$, which is designed to preserve performance while applying masks to the input image. As β increases, we observe a gradual increase in both accuracy and AUC, accompanied by a decrease in fairness metrics (EO, DP, and ADF). This trade-off occurs because a higher β value allows the generator to focus more on preserving diagnostic performance, but at the cost of reduced fairness, as the generator becomes less constrained in limiting sensitive attribute influence.

5. Conclusion

This work introduces Diagnosis Fairness (DF) to mitigate demographic biases in medical imaging, along with a new metric, Approximate Diagnosis Fairness (ADF), for practical fairness evaluation. Our approach, combining data utility enhancement and adversarial training, improves fairness across medical datasets without sacrificing diagnostic accuracy. Experimental results show that our method effectively reduces bias while maintaining robust and interpretable performance, underscoring the potential of causal reasoning in fair medical AI. Future work will focus on more efficient causal modeling for high-dimensional data and addressing the balance between fairness and accuracy, especially for subgroups where strict fairness constraints may impact performance.

Appendix A. Implementation details

A.1. Hyperparameter Settings

In our experiments, we explored several key hyperparameters to evaluate their impact on performance and fairness. Specifically, the noise strength η , controlling the intensity of adversarial perturbations generated by the model’s generator, is set to 0.2 in experiments. The fairness weighting parameter α is set to 1.0, balancing the entropy-based fairness loss with task performance. Loss regularization β is set to 1.0 to prioritize classification accuracy. The learning rate for the generator, discriminator, and feature extractor was initialized at 1×10^{-4} , with step-based learning rate schedulers applied to ensure convergence. Training and testing datasets were split with a ratio of 90% to 10%, ensuring sufficient data for both robust evaluation and training.

A.2. Dataset-Specific Tasks

Regarding dataset labels, for the MIMIC-CXR^[30] and CheXpert^[31] datasets, our main task focused on pneumonia classification, utilizing gender as the sensitive attribute (mapped as male = 0, female = 1). For the TCGA^[32] dataset, we focused on detecting pathological stages of lung cancer, specifically distinguishing between early and late stages. We preprocess the dataset to ensure a balanced representation of these stages (Stage I A, Stage I B, Stage II A, and Stage II B are mapped to the early stage; Stage III A, Stage III B, Stage III C, Stage IV A, and Stage IV B are mapped to the late stage), enabling robust evaluation of both performance and fairness metrics in the classification task. Also, gender is utilized as the sensitive attribute.

A.3. Environments

The models are trained offline using PyTorch^[35] and executed on a machine equipped with an AMD EPYC 7763 64-Core Processor CPU @ 4.00GHz and an NVIDIA RTX 6000 Ada Generation GPU, running the Ubuntu 22.04.3 LTS operating system. The experiments were conducted within a Conda environment and a Docker container to ensure reproducibility and ease of deployment. The Conda environment was managed using Miniconda^[36] version 23.9.0 (Python 3.10.13), while the Docker container^[37] was built on Docker version 24.0.5 with a base image of “nvidia/cuda:12.4.0-cudnn8-devel-ubuntu22.04” to support GPU acceleration. We will provide our Conda environment, Docker container, and code implementations upon publication.

References

1. [^]Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JAWM, Van Ginneken B, Sánchez CI. "A survey on deep learning in medical image analysis". *Medical image analysis*. **42**: 60–88, 2017.
2. [^]Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, Kim N (2017). "Deep learning in medical imaging: general overview". *Korean journal of radiology*. **18** (4): 570–584.
3. [^]Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine*. **15** (11): e1002686, 2018.

4. [^]Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". *Science*. 366 (6464): 447–453. doi:[10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342).
5. [^]Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH (2018). "Ensuring Fairness in Machine Learning to Advance Health Equity". *Annals of Internal Medicine*. 169 (12): 866–872. doi:[10.7326/M18-1990](https://doi.org/10.7326/M18-1990).
6. [^]Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L-C, Correa R, Dullerud N, Ghassemi M, Huang S-C, Kuo P-C, Lungren MP, Palmer LJ, Price BJ, Purkayastha S, Pyrros AT, Oakden-Rayner L, Okechukwu C, Seyyed-Kalantari L, Trivedi H, Wang R, Zaiman Z, Zhang H (2022). "AI recognition of patient race in medical imaging: a modeling study". *The Lancet Digital Health*. 4 (6): e406–e414. doi:[10.1016/s2589-7500\(22\)00063-2](https://doi.org/10.1016/s2589-7500(22)00063-2).
7. [^]Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E (2020). "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis". *Proceedings of the National Academy of Sciences*. 117 (23): 12592–12594. doi:[10.1073/pnas.1919012117](https://doi.org/10.1073/pnas.1919012117).
8. [^]Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M (2020). "CheXclusion: Fairness gaps in deep chest X-ray classifiers". *arXiv*. Available from: <https://arxiv.org/abs/2003.00827>.
9. [^]Chen Z, Zhang JM, Sarro F, Harman M (2024). "Fairness Improvement with Multiple Protected Attributes: How Far Are We?" *arXiv*. [arXiv:2308.01923 \[cs.LG\]](https://arxiv.org/abs/2308.01923).
10. ^{a, b, c, d, e, f, g, h}Zhang BH, Lemoine B, Mitchell M (2018). "Mitigating Unwanted Biases with Adversarial Learning." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. ACM. p. 335–340. doi:[10.1145/3278721.3278779](https://doi.org/10.1145/3278721.3278779).
11. [^]Madras D, Creager E, Pitassi T, Zemel R (2018). "Learning Adversarially Fair and Transferable Representations". *arXiv*. Available from: <https://arxiv.org/abs/1802.06309>.
12. [^]Alvi M, Zisserman A, Nellerker C. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: *Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8–14, 2018, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag; 2019. p. 556–572. doi:[10.1007/978-3-030-11009-3_34](https://doi.org/10.1007/978-3-030-11009-3_34).
13. [^]Burlina P, Joshi N, Paul W, Pacheco KD, Bressler NM (2020). "Addressing Artificial Intelligence Bias in Retinal Disease Diagnostics". *arXiv*. Available from: <https://arxiv.org/abs/2004.13515>.
14. ^{a, b}Paul W, Hadzic A, Joshi N, Alajaji F, Burlina P (2021). "TARA: Training and Representation Alteration for AI Fairness and Domain Generalization". *arXiv*. [arXiv:2012.06387 \[cs.LG\]](https://arxiv.org/abs/2012.06387).
15. [^]Cheng V, Suriyakumar VM, Dullerud N, Joshi S, Ghassemi M. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In: *Proceedings of the 202*

- 1 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2021. p. 149–160. doi:~~10.1145/3442188.3445879~~.
16. ^aZhou Y, Huang SC, Fries JA, Youssef A, Amrhein TJ, Chang M, Banerjee I, Rubin D, Xing L, Shah N, Lungren MP (2021). "RadFusion: Benchmarking Performance and Fairness for Multimodal Pulmonary Embolism Detection from CT and EHR". arXiv. [arXiv:2111.11665](https://arxiv.org/abs/2111.11665) [eess.IV].
17. ^a, ^b, ^c, ^d, ^e, ^f, ^gWang Z, Dong X, Xue H, Zhang Z, Chiu W, Wei T, Ren K. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 10379–10388.
18. ^a, ^bPearl J, et al. Models, reasoning and inference. Cambridge, UK: Cambridge University Press. 19 (2): 3, 2000.
19. ^a, ^b, ^cZhang L, Wu Y, Wu X (2016). "A causal framework for discovering and removing direct and indirect discrimination". arXiv preprint arXiv:1611.07509. Available from: <https://arxiv.org/abs/1611.07509>.
20. ^a, ^b, ^cPearl J (2022). "Direct and indirect effects." In: Probabilistic and causal inference: the works of Judea Pearl, pp. 373–392.
21. ^aPearl J (2009). "Causal inference in statistics: An overview". Statistics surveys 3. pages 96–146.
22. ^a, ^bHardt M, Price E, Srebro N (2016). "Equality of opportunity in supervised learning". Advances in neural information processing systems. 29.
23. ^a, ^bDwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012). "Fairness through awareness." In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214--226.
24. ^aShannon CE (1948). "A mathematical theory of communication". The Bell system technical journal. 27 (3): 379–423.
25. ^aHjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y (2019). "Learning deep representations by mutual information estimation and maximization." In: International Conference on Learning Representations (ICLR). Available from: <https://arxiv.org/abs/1808.06670>.
26. ^aNowozin S, Cseke B, Tomioka R. "f-GAN: Training generative neural samplers using variational divergence minimization." In: Advances in Neural Information Processing Systems (NeurIPS). 2016. p. 271–279. Available from: <https://arxiv.org/abs/1606.00709>.
27. ^aLi A, Duan Y, Yang H, Chen Y, Yang J (2020). "TIPRDC: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations". In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020: 824–832.

28. [△]Bayes T (1763). "An Essay Towards Solving a Problem in the Doctrine of Chances". *Philosophical Transactions of the Royal Society of London*. 53: 370–418. doi:[10.1098/rstl.1763.0053](https://doi.org/10.1098/rstl.1763.0053).
29. [△]Grimmett G, Stirzaker D. *Probability and random processes*. Oxford university press; 2020.
30. [△], [△]Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng C, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S (2019). "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs". *arXiv preprint arXiv:1901.07042*. [arXiv:1901.07042](https://arxiv.org/abs/1901.07042).
31. [△], [△]Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghighi B, Ball R, Shpanskaya K, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence*. 2019; 33(01):590–597.
32. [△], [△]Albertina B, Watson M, Holback C, Jarosz R, Kirk S, Lee Y, Rieger-Christ K, Lemmerrman J (2016). "The cancer genome atlas lung adenocarcinoma collection (tcga-luad)(version 4)[data set]". *The Cancer Imaging Archive*.
33. [△], [△], [△], [△], [△]He K, Zhang X, Ren S, Sun J. "Deep Residual Learning for Image Recognition." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016: 770–778. doi:[10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
34. [△], [△]Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; 2017. p. 618–626.
35. [△]Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". *Advances in neural information processing systems*. 32.
36. [△]Anaconda, Inc. (2023). Miniconda: A free minimal installer for conda [software]. Available from: <https://docs.conda.io/en/latest/miniconda.html>.
37. [△]Docker, Inc. Docker: A platform to build, run, and share applications with containers [software]. 2024. Version 24.0.5. Available from: <https://www.docker.com/>.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.