nature communications

Article

https://doi.org/10.1038/s41467-024-52930-1

Enhancing fairness in AI-enabled medical systems with the attribute neutral framework

Received: 6 November 2023

Accepted: 25 September 2024

Published online: 10 October 2024

Check for updates

Lianting Hu[1,2,3,4,10], Dantong Li[2,3,4,10], Huazhang Liu[3,4], Xuanhui Chen[3,4], Yunfei Gao[3,4], Shuai Huang[3,4], Xiaoting Peng[3,4], Xueli Zhang[4,5,6], Xiaohe Bai[7], Huan Yang[2,3,4], Lingcong Kong[3,4], Jiajie Tang[8], Peixin Lu[9], Chao Xiong[1] ✉ & Huiying Liang[1,2,3,4] ✉

Questions of unfairness and inequity pose critical challenges to the successful deployment of artificial intelligence (AI) in healthcare settings. In AI models, unequal performance across protected groups may be partially attributable to the learning of spurious or otherwise undesirable correlations between sensitive attributes and disease-related information. Here, we introduce the Attribute Neutral Framework, designed to disentangle biased attributes from disease-relevant information and subsequently neutralize them to improve representation across diverse subgroups. Within the framework, we develop the Attribute Neutralizer (AttrNzr) to generate neutralized data, for which protected attributes can no longer be easily predicted by humans or by machine learning classifiers. We then utilize these data to train the disease diagnosis model (DDM). Comparative analysis with other unfairness mitigation algorithms demonstrates that AttrNzr outperforms in reducing the unfairness of the DDM while maintaining DDM's overall disease diagnosis performance. Furthermore, AttrNzr supports the simultaneous neutralization of multiple attributes and demonstrates utility even when applied solely during the training phase, without being used in the test phase. Moreover, instead of introducing additional constraints to the DDM, the AttrNzr directly addresses a root cause of unfairness, providing a model-independent solution. Our results with AttrNzr highlight the potential of data-centered and model-independent solutions for fairness challenges in AI-enabled medical systems.

Artificial Intelligence (AI) technology has made tremendous progress in recent years, and its applications in the medical field are increasing[1–3]. While AI has achieved specialist-level performance, its direct application has raised concerns about producing unfair outcomes across various scenarios[4–7]. Instances such as Gichoya et al.'s study on AI systems capable of detecting a patient's race with differing degrees of accuracy across self-reported racial groups in medical imaging[8], and Seyyed-Kalantari et al.'s findings on underdiagnosis across age, sex, race, and socioeconomic status, underscore the critical issue of fairness[9]. Unfairness, characterized by uneven performance among groups identified by sensitive attributes, is often the result of AI-enabled medical systems relying on improper correlations stemming from attribute biases[5,10].

Previous efforts to mitigate these biases have involved specific adjustments to individual models[11,12]. Puyol-Antón et al. proposed a fairness meta-learning for segmentation, in which a deep learning

A full list of affiliations appears at the end of the paper. ✉e-mail: xcpopo11@yeah.net; lianghuiying@hotmail.com

classifier is trained to classify race and jointly optimized with the segmentation model[13]. Dash et al. designed a counterfactual regularizer to mitigate the bias of a pre-trained machine learning classifier[14]. Although effective, such model-specific modifications demand substantial resources and expert knowledge for effective bias identification and correction, a task further complicated by the scarcity of domain experts. Hence, a model-independent solution is increasingly recognized as the most effective strategy.

Commonly proposed approaches to mitigate AI model unfairness act debiasing on the feature encoding space[11] or are based on some kind of image augmentation strategy[12]. Yet, several studies have indicated that such biases are closely linked to both disease outcomes and the features of medical images. Effectively separating and eliminating biased attributes while preserving essential medical information on the image level remains a substantial challenge, with potential solutions still under exploration.

To address this challenge, we propose the Attribute Neutral Framework. This framework aims to disentangle biased attributes from disease-relevant information at the data level and subsequently neutralize them to guarantee a balanced representation across diverse subgroups. Through the simultaneous neutralization of multiple attributes, the Attribute Neutral Framework tackles biased attributes at their root, offering a universally applicable solution that transcends model complexities and deployment scenarios. Within the framework, we develop the Attribute Neutralizer (AttrNzr) to generate neutralized data, for which protected attributes can no longer be easily recognized by humans or by machine learning classifiers. Then, we utilize these data to train the disease diagnosis model (DDM). Comparative analysis with other unfairness mitigation algorithms demonstrates that AttrNzr effectively reduces the unfairness of the DDM while maintaining its overall disease diagnosis performance. The primary contributions of this article are twofold: 1) Exploring the effectiveness of using neutralized data to mitigate the unfairness of AI-enabled medical systems and employing AttrNzr to achieve neutralization of X-ray images across multiple attributes; 2) Offering a model-independent solution by directly addressing unfairness at the image level, thus obviating the need for individual modifications to each model.

Results

The study verified the mitigation effect of the AttrNzr on the unfairness within AI-enabled medical systems using three large public chest X-ray image datasets: ChestX-ray14[15], MIMIC-CXR[16], and CheXpert[17]. The metadata of the datasets includes the attributes of sex and age. In the MIMIC-CXR dataset, the additional attributes of race and insurance are available. The AttrNzr was trained using these datasets to modify the attribute intensity within X-ray images. The modification intensity α controls the degree of attribute modification in the AttrNzr. α ranges from 0 to 1, with 0 indicating no modification, 1 indicating negation of the attribute, and 0.5 indicating a neutral attribute. The attribute recognition, involving both AI judges and human judges, was conducted to assess the ability of the AttrNzr to generate X-ray images with specific attributes, which are indistinguishable from genuine X-ray images possessing those attributes. Multiple DDMs were built, including those based on neutralized X-ray images (modification intensities: 0.5), modified X-ray images (modification intensities: 0.6, and 0.7), and those integrating other unfairness mitigation algorithms: the Fairmixup[12], the Fairgrad[11], and the Balanced sampling[18,19]. To assess the unfairness within the DDM, three types of unfairness metrics: worst-case performance, performance gap, and performance standard deviation are employed. Lastly, to mitigate the unfairness of the preexisting AI-enabled medical system and to reduce the computational requirements of the AttrNzr at the test stage, two application paradigms of the AttrNzr: test-stage neutralization and training-stage neutralization were proposed. An overview of our comprehensive study is presented in Fig. 1.

X-ray images generated by attribute neutralizers

Based on the original X-ray images and their corresponding attribute labels, AttrNzrs for single and multiple attributes were trained. These AttrNzrs were used to modify the attributes with a modification intensity α, which controls the degree of attribute modification. α ranges from 0 to 1, with 0 indicating no modification, 1 indicating negation of the attribute, and 0.5 indicating a neutral attribute. The average modified X-ray images and some examples are shown in Fig. 2.

The Structural Similarity Index Measure (SSIM) was employed to quantitatively evaluate the similarity between the modified X-ray images and the original X-ray image. The results indicate a gradual decrease in SSIM as the modification intensity increases. As shown in Fig. 2a–d, there is a strong negative correlation between the SSIM and the modification intensity in these average modified X-ray images (Pearson's r of (male to female): −0.9097, Pearson's r of (female to male): −0.8917, Pearson's r of (≥60 y male to <60 y female): −0.9099, Pearson's r of (<60 y female to ≥60 y male): −0.9040). Additionally, the difference image produced by the AttrNzr is similar to the difference image produced by original X-ray images (Supplementary Fig. S6).

The aforementioned observations in average X-ray images are also observable in the modified X-ray examples. Figure 2e–h showcases examples of the modified X-ray images. Visually, as the modification intensity increases, the modified X-ray image undergoes subtle changes gradually (Supplementary Movie 1). The difference image illustrates that higher modification intensities lead to increased differences between the original X-ray images and the modified ones. Moreover, the difference regions observed align with anatomical differences between different subgroups, such as differences between male and female mammary glands and differences in skeletal size between older (≥60 y) and younger (<60 y) age groups.

Attributes recognition of modified X-ray images

In this study, attribute recognition was used to assess the ability of the AttrNzr to generate X-ray images with specific attributes, which were indistinguishable from real X-ray images with those attributes. In the first attribute recognition, AI judges were fully trained on original X-ray images and then asked to identify the sex (female/male) and age (<60 y/ ≥ 60 y) of modified X-ray images.

As shown in Fig. 3a–d, for modification intensities below 0.5, the AI judge exhibits proficient performance in accurately identifying the original attributes of the modified X-ray images, demonstrating a reasonable agreement between the predicted and original attributes. Conversely, for modification intensities above 0.5, it is difficult for the AI judges to correctly identify the original attributes of the modified X-rays. Since sex and age are binary, the attributes predicted by the AI judge are almost opposite to the real attributes. Consequently, the modification intensity of 0.5 is the turning point of the AI judge's identification performance. As depicted in Fig. 3g, h, while the modification intensity is set at 0.5, the distance between attribute groups reaches its minimum. As the modification intensity deviates further from 0.5, the distance between attribute groupings gradually expands. This observation is also substantiated by the UMAP visualizations of attribute groups, as shown in Supplementary Figs. S7 and S8. Figure 3i–l further depicted that modified X-ray images possessing specific attributes were more likely to activate the corresponding attribute gradient. Additionally, notable dissimilarities emerged between the activation regions associated with modification intensities greater than 0.5 and those below 0.5, highlighting the significance of 0.5 as the turning point for changes in the activation heatmap.

In the second attribute recognition, human judges were asked to identify the attributes of X-ray images generated by our AttrNzr. Figure 3e, f demonstrate a strong negative correlation between the accuracy of attribute identification by human judges and the modification intensity (Pearson's r of age: −0.9919, Pearson's r of sex:

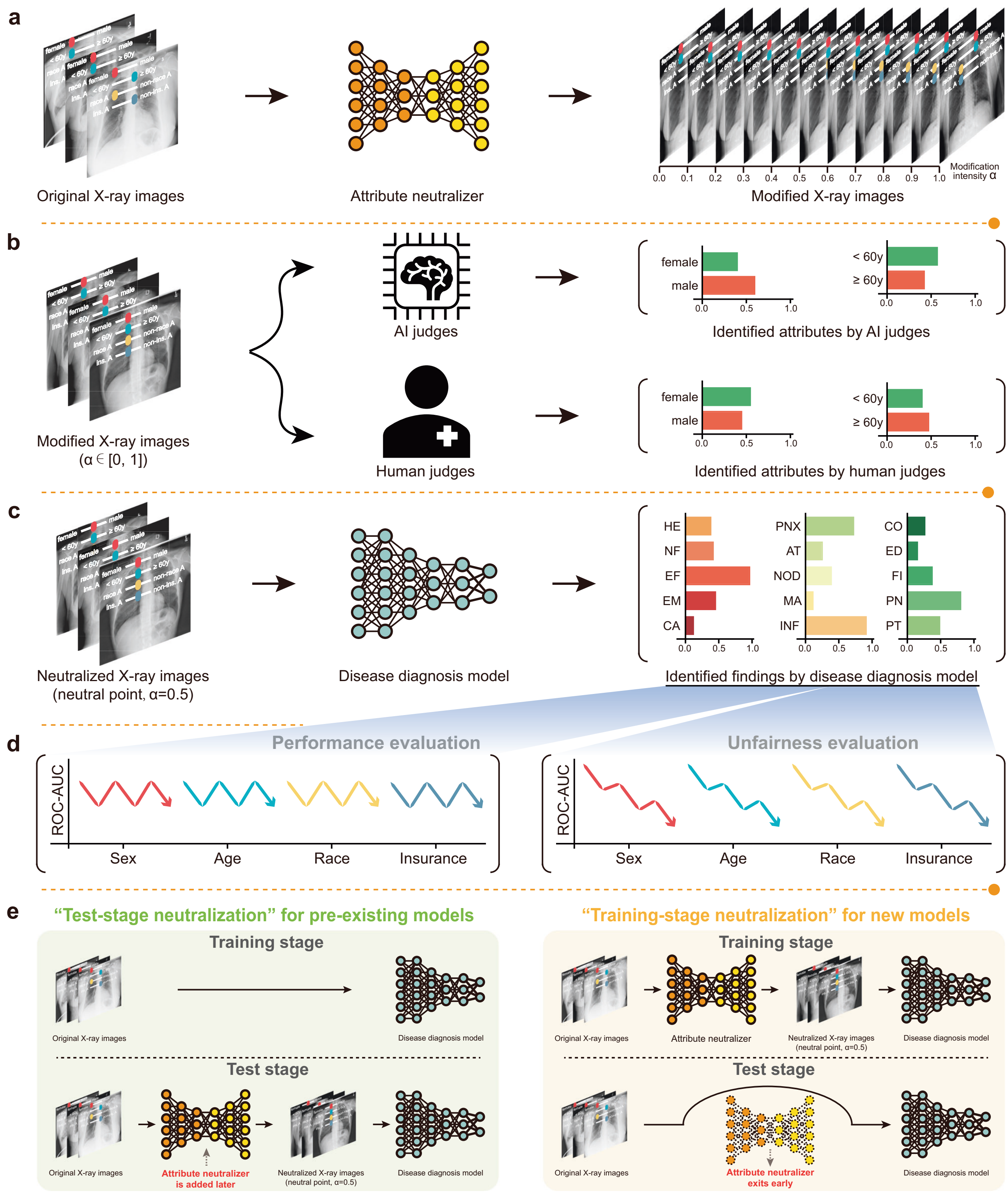

Fig. 1 | A comprehensive framework of our study. a Utilizing the AttrNzr to create modified X-ray images, where the attributes of these modified images can be regulated by the modification intensity α. **b** Involving AI judges and human judges in the attribute recognition to discriminate the original attributes of the modified X-ray images. **c** Establish a DDM on neutralized X-ray images generated by the AttrNzr with an α value of 0.5. **d** Conduct performance and unfairness evaluations based on predictions from the DDM. **e** Investigate the implementation of AttrNzrs in both the test-stage neutralization paradigm to mitigate unfairness in pre-existing medical AI models, as well as the training-stage neutralization paradigm to minimize computational requirements in new medical AI models.

−0.8896). Notably, human judges exhibited superior proficiency in identifying the sex compared to the age in X-ray images. Furthermore, the identification performance of human judges changes more smoothly than that of AI judges.

The identification results of both AI judges and human judges indicated that the X-ray images produced by AttrNzr achieved a high degree of authenticity. Furthermore, the modification intensity adeptly regulates the strength of various subgroup attributes. Notably,

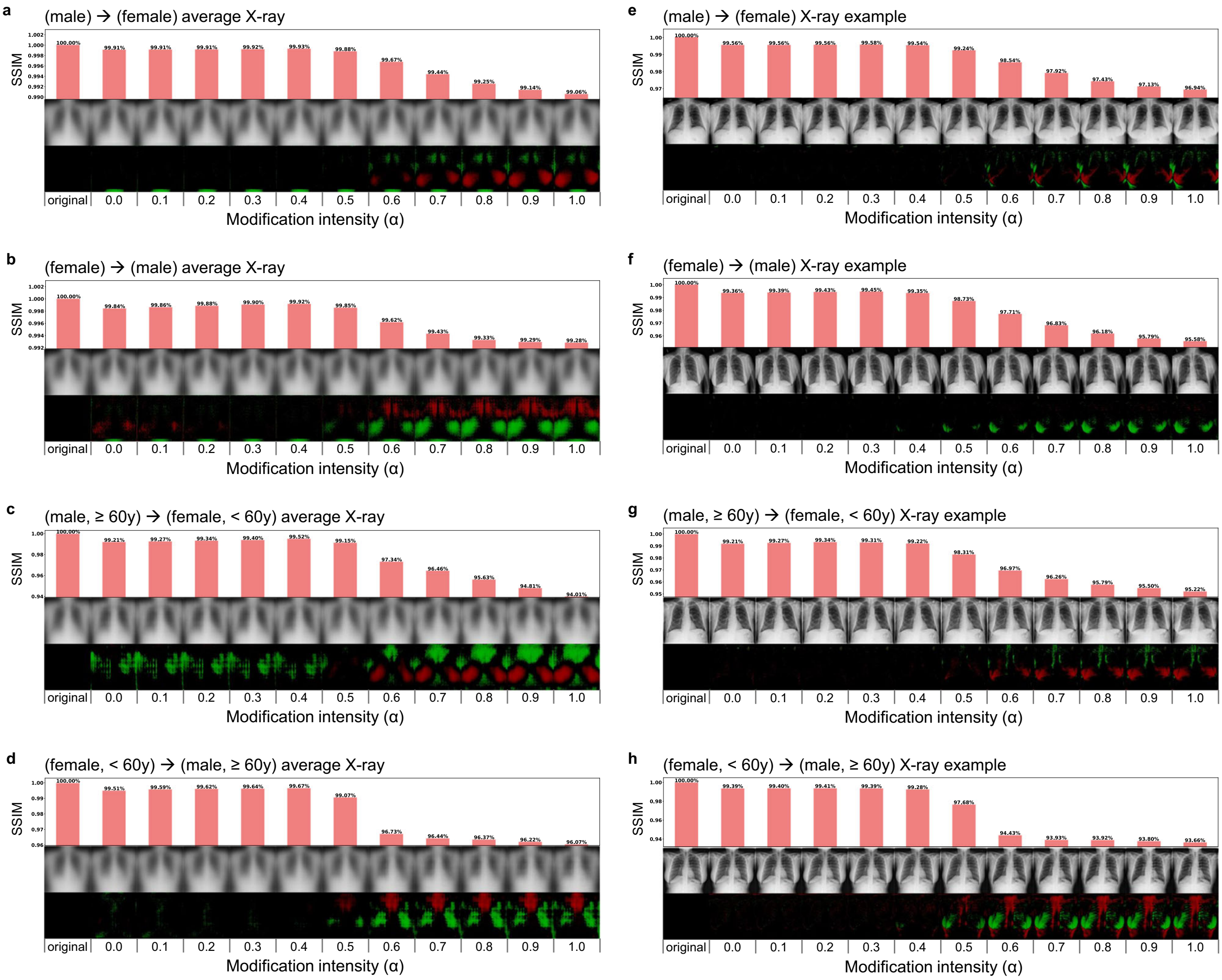

Fig. 2 | Modified X-ray images generated by AttrNzrs. a–d Average modified X-ray images of the attribute modification. The modified attributes are (male) to (female), (female) to (male), (male, ≥60 y) to (female, <60 y), and (female, <60 y) to (male, ≥60 y), respectively. Other average modified X-ray images with other modified attributes are shown in Supplementary Fig. S4. **e–h** Examples of modified X-ray images. The modified attributes are the same as **a**–**d**. Examples of X-ray images with other modified attributes are shown in Supplementary Fig. S5. Each subfigure consists of four rows. The first row displays a bar chart depicting the Structural Similarity Index Measure (SSIM) between the modified X-ray image and the original X-ray image. The second row shows the original X-ray image and the modified X-ray images. The third row displays the difference image between the modified X-ray images and the original X-ray image. In comparison to the original X-ray image, regions of high intensity in the modified X-ray image are marked in red, while regions of low intensity are marked in green. The fourth row exhibits the modification intensities α with values ranging from 0.0 to 1.0 sequentially (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0). **a**, **b**, **e**, and **f** are generated by the single-attribute AttrNzrs, while **c**, **d**, **g**, and **h** are generated by the multi-attribute AttrNzrs.

both AI judges and human judges exhibited the greatest uncertainty when confronted with a modification intensity of 0.5. This observation underscores that the modified ($\alpha = 0.5$) X-ray image, unlike the binary attributes at their extreme ends, represents a form of neutral and unbiased data.

Disease diagnosis performance

Is the disease-related information in the X-ray image retained during the attribute modification process by the AttrNzr? To address this question, four DDMs were trained on the original and modified X-ray images, respectively. The modifications pertained to the attribute of age, with intensities of modification (α) set at 0.5, 0.6, and 0.7. Importantly, at $\alpha = 0.5$, the X-ray images generated by the AttrNzr exhibit signs of attribute neutralization, hence referred to as "neutralized X-ray images." Additionally, alternative unfairness mitigation algorithms, namely, the Fairmixup[12], the Fairmixup manifold[12], the Fairgrad[11], and the Balanced sampling[18] were also applied to train another four DDMs.

Figure 4a–d and Supplementary Fig. S9 depict 7 Critical Difference (CD) diagrams of various unfairness mitigation algorithms in disease diagnosis. These 7 diagrams correspond to 7 metrics: Macro-(ROC-AUC, accuracy, sensitivity, specificity, precision, F1-score, and PR-AUC). It is evident that the overall performance of the 7 DDMs across the 6 {datasets, attributes} combinations exhibit significant differences (Friedman test *P*-value: <0.05 across all 7 metrics). The rankings of original-based DDM on the 7 metrics are 1.5, 3.3, 2.8, 3.3, 2.2, 2.7, and 1.3. Conversely, the rankings of neutralized-based DDM on the 7 metrics are 3.2, 1.0, 6.5, 1.0, 1.0, 1.0, and 2.5, respectively. Notably, the Macro-sensitivity of the neutralized-based DDM is significantly lower than that of the original-based DDM, with no significant differences observed in the other 6 metrics. Regarding other unfairness mitigation algorithms, the Fairgrad-based DDM exhibits significantly lower performance than the original-based DDM in terms of Macro-(ROC-AUC, precision, F1-score, and PR-AUC). Both the Fairmixup-based and the Fairmixup-manifold-based DDMs show significantly lower performance compared to the original-based DDM concerning

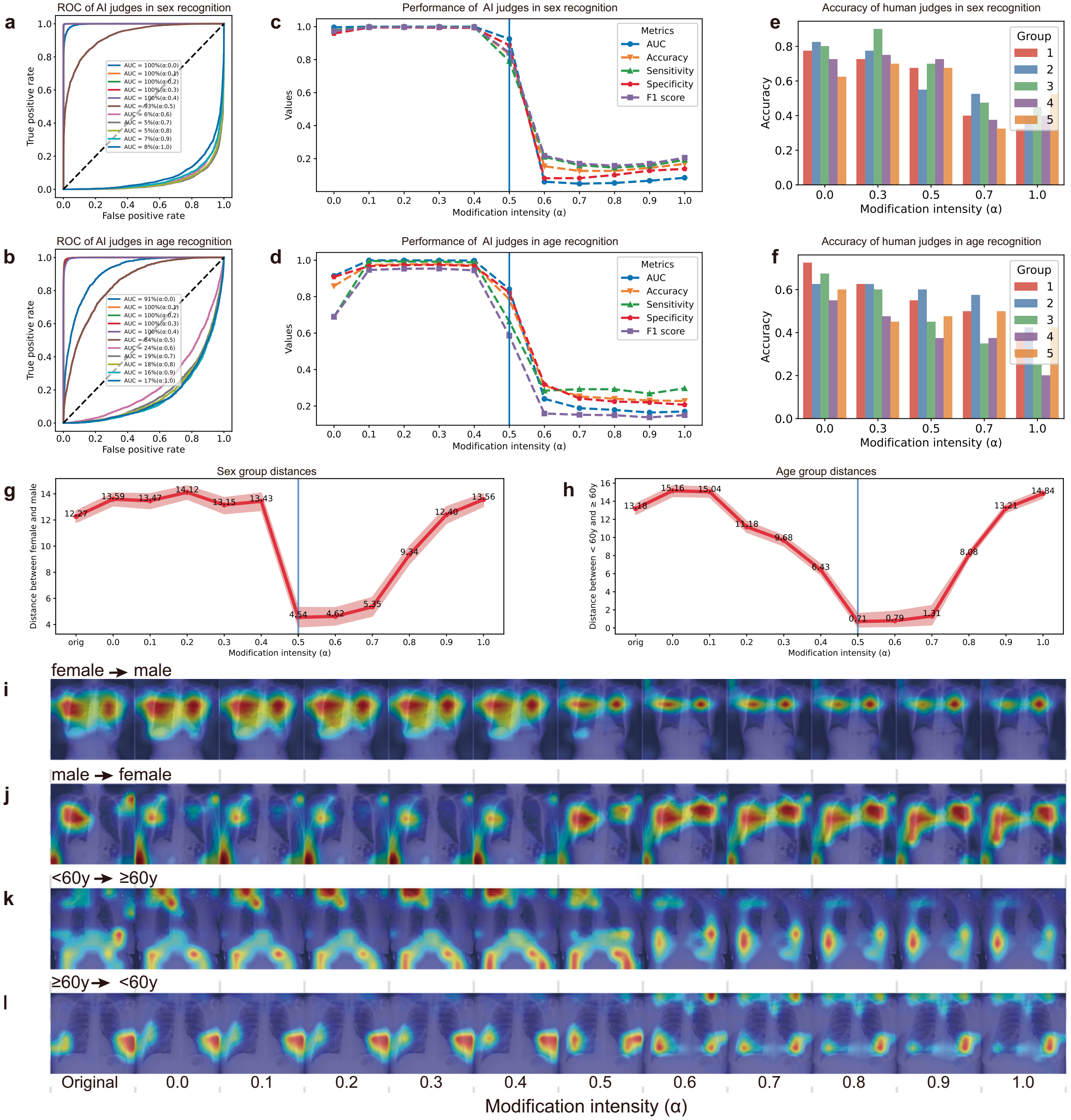

Fig. 3 | Performance of AI judges and human judges in attribute recognition. Receiver operating characteristic (ROC) curves of AI judges in attribute recognition for sex (**a**) and age (**b**). The AI judges were asked to identify the attributes of modified X-ray images. The modification intensity α was set to 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. Area under the curve (AUC), accuracy, sensitivity, specificity, and F1 score of AI judges in attribute recognition for sex (**c**) and age (**d**) at different modification intensities. Accuracy of human judges in the attribute recognition for sex (**e**) and age (**f**). Human judges were asked to identify the attributes of five groups of X-ray images, modified at intensities of 0.0, 0.3, 0.5, 0.7, and 1.0. The sex (**g**) and age (**h**) group distances, computed using 2-dimensional features obtained through UMAP reduction from the AI judge's 768-dimensional final layer features. Distances are measured as Euclidean distances between group centers, with a 95% confidence interval derived from non-parametric bootstrapping (1000 iterations). The visualization of X-ray images after UMAP dimension reduction is shown in Supplementary Fig. S7 (sex) and Supplementary Fig. S8 (age). **i**, **j** Two activation heatmap examples of AI judges in the attribute recognition for sex. The activation category of the heatmap is female. The original sex of the examples is female (**i**) and male (**j**). **k**, **l** Two activation heatmap examples of AI judges in the attribute recognition for age. The activation category of the heatmap is <60 y. The original age group of the examples is <60 y (**k**) and ≥60 y (**l**). Source data are provided as a Source Data file.

Macro-(ROC-AUC, precision, and PR-AUC). However, Modi (α = 0.6)-based, and Balanced-sampling-based DDMs have no significant difference in all 7 metrics, compared to the original-based DDM. For more detailed Macro-performance information, please refer to Supplementary Table S9–11.

In the comparative analysis of DDMs, ROC curves, and PR curves are utilized to evaluate specific findings. A total of 43 findings are considered, distributed as follows: 15 in ChestX-ray14, 14 in MIMIC-CXR, and 14 in CheXpert datasets. Delong's test and the Bootstrap method are employed to assess the statistical significance of

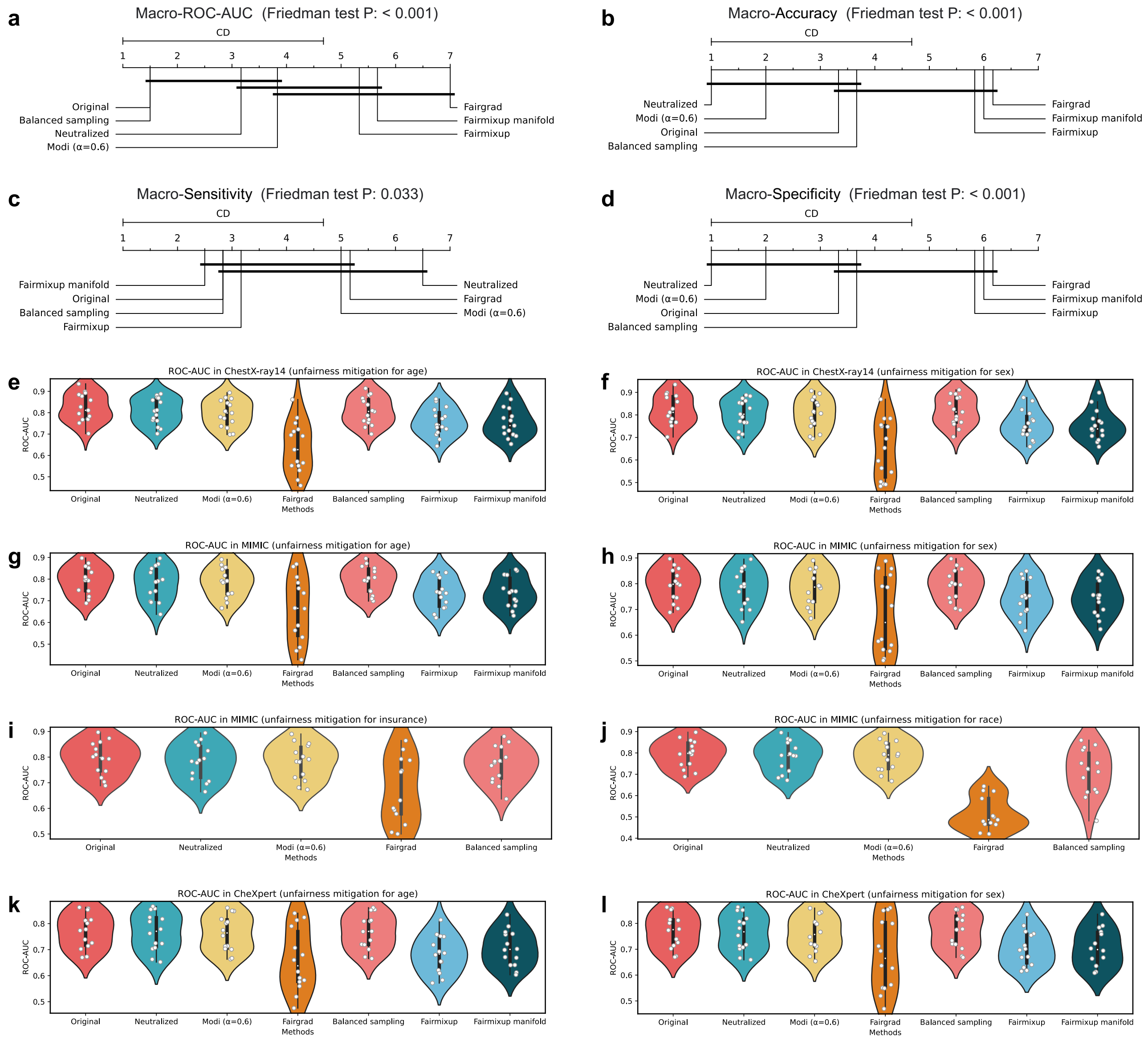

Fig. 4 | Performance comparison of various unfairness mitigation algorithms in disease diagnosis. Critical Difference (CD) diagrams for Macro-ROC-AUC (**a**), Macro-Accuracy (**b**), Macro-Sensitivity (**c**), and Macro-Specificity (**d**). In each diagram, the Friedman test and the Nemenyi post-hoc test are performed across 6 {dataset, attribute} combinations: {ChestX-ray14, age}, {ChestX-ray14, sex}, {MIMIC-CXR, age}, {MIMIC-CXR, sex}, {CheXpert, age}, and {CheXpert, sex}. The CD value is 3.68. Violin plots of ROC-AUC for various DDMs in ChestX-ray14 (**e**, **f**), MIMIC-CXR (**g**–**j**), and CheXpert (**k**, **l**). The violin plot shows the distribution of ROC-AUCs across all findings (15 findings in ChestX-ray14, 14 findings in MIMIC-CXR, and 14 findings in CheXpert). The attributes corresponding to unfairness mitigation include age (**e**, **g**, **k**), sex (**f**, **h**, **l**), insurance (**i**), and race (**j**). In the violin plot, the central white dot represents the median, while the thick line inside the violin indicates the interquartile range. The whiskers represent the range of the data, excluding outliers. Source data are provided as a Source Data file.

differences in ROC curves and PR curves, respectively. For ROC curves (Supplementary Fig. S10 for examples), when compared to the original-based DDM (Supplementary Tables S12–14), the counts of findings exhibiting no significant difference are as follows: 15 (neutralized), 13 (Modi ($\alpha = 0.6$)), 11 (Modi ($\alpha = 0.7$)), 1 (Fairgrad), 28 (Balanced sampling), 0 (Fairmixup), and 0 (Fairmixup manifold). Additionally, within the neutralized-based DDM, the ROC-AUCs of all findings demonstrate an even distribution (Fig. 4e–l). Regarding PR curves (Supplementary Fig. S11 for examples), when compared to the original-based DDM (Supplementary Table S15–17), the counts of findings with no significant difference are 23 (neutralized), 17 (Modi ($\alpha = 0.6$)), 13 (Modi ($\alpha = 0.7$)), 0 (Fairgrad), 20 (Balanced sampling), 0 (Fairmixup), and 0 (Fairmixup manifold). Moreover, within the neutralized-based DDM, the PR-AUCs of all findings exhibit an even distribution (Supplementary Fig. S12).

Figure 5 showcases four examples evaluated by AI judges, human judges, and DDMs. It can be observed that the neutralized X-ray image increases the uncertainty of the AI judges and human judges in identifying the attributes compared to the original X-ray image. However, the neutralized-based DDM can still identify the corresponding findings from the neutralized X-ray images, and the detection results are consistent with those of the original-based DDM.

Unfairness of disease diagnosis models

This section aims to assess the unfairness of various DDMs using three types of unfairness metrics: worst-case performance among subgroups[20,21], the performance gap between the best and worst subgroups[20], and the performance standard deviation across all subgroups[13,22], is introduced. Within the unfairness evaluation,

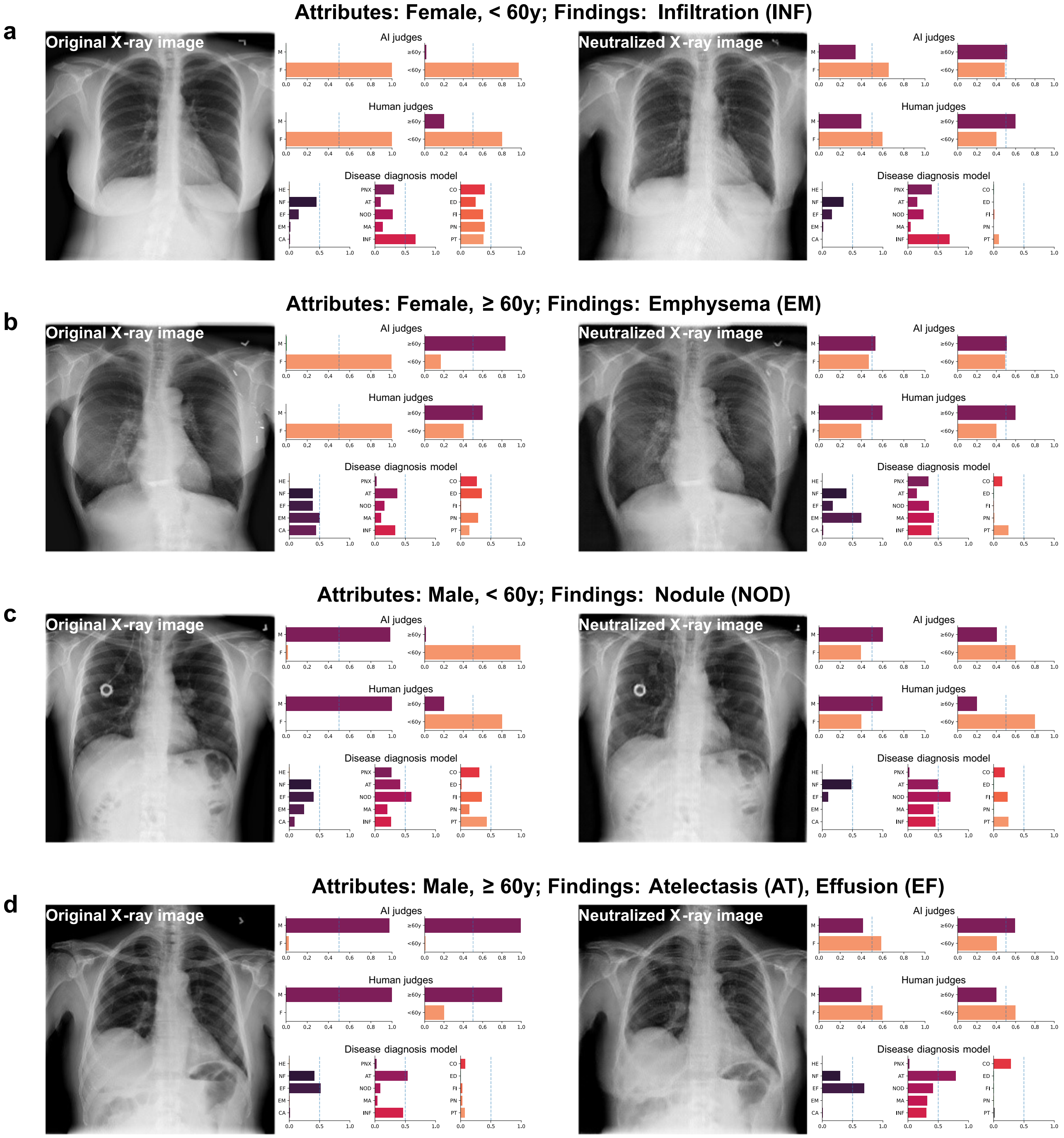
a
Attributes: Female, < 60y; Findings: Infiltration (INF)
b
Attributes: Female, ≥ 60y; Findings: Emphysema (EM)
c
Attributes: Male, < 60y; Findings: Nodule (NOD)
d
Attributes: Male, ≥ 60y; Findings: Atelectasis (AT), Effusion (EF)
Original X-ray image
Neutralized X-ray image
AI judges
Human judges
Disease diagnosis model
M
F
≥60y
<60y
HE
NF
EF
EM
CA
PNX
AT
NOD
MA
INF
CO
ED
FI
PN
PT
0.0
0.2
0.4
0.6
0.8
1.0
0.5

Fig. 5 | Four examples showcasing disease diagnosis on original and neutralized X-ray images. The attributes for each example are as follows: **a** female, <60 y; **b** female, ≥60 y; **c** male, <60 y; and **d** male, ≥60 y. The identified findings for these examples are as follows: **a** infiltration; **b** emphysema; **c** nodule; and **d** atelectasis, effusion. Each subfigure consists of an original X-ray image and its corresponding neutralized X-ray image. The neutralized attributes are sex and age, and the modification intensity α is 0.5. The attributes of the X-ray images are identified by AI judges and human judges, while the findings of X-ray images are identified by the DDM. AI judges and DDMs report output probabilities. Human judges report the voting ratio based on the evaluations of five human judges. AT Atelectasis, CA Cardiomegaly, CO Consolidation, ED Edema, EF Effusion (EF), EM Emphysema, FI Fibrosis, HE Hernia, INF Infiltration, MA Mass, NOD Nodule, PT Pleural Thickening, PN Pneumonia, PNX Pneumothorax, NF No Finding.

performance is evaluated based on ROC-AUC[20], accuracy[19,23,24], sensitivity[23], and specificity[20].

Figure 6a–d and Supplementary Fig. S13 present unfairness CD diagrams illustrating various unfairness mitigation algorithms. It is evident that the overall unfairness of the 7 DDMs across the 6 {datasets, attributes} combinations exhibit significant differences (Friedman test P-value: <0.05 across 11 metrics except for worst-case sensitivity). The rankings of neutralized-based DDM on the 12 unfairness metrics are 1.3, 1.8, 2.2, 1.8, 1.0, 1.0, 2.7, 1.0, 1.0, 1.8, 2.2, 1.8, and ranks first in performance-SD, worst-case-performance. The rankings of Fairgrad-based DDM on the 12 unfairness metrics are 3.8, 1.2, 1.0, 1.2, 7.0, 6.5, 5.0, 6.3, 7.0, 1.2, 1.0, 1.2, and ranks first in the accuracy, sensitivity, and specificity gap. Furthermore, there is no significant difference between the unfairness of balanced-sampling-based and

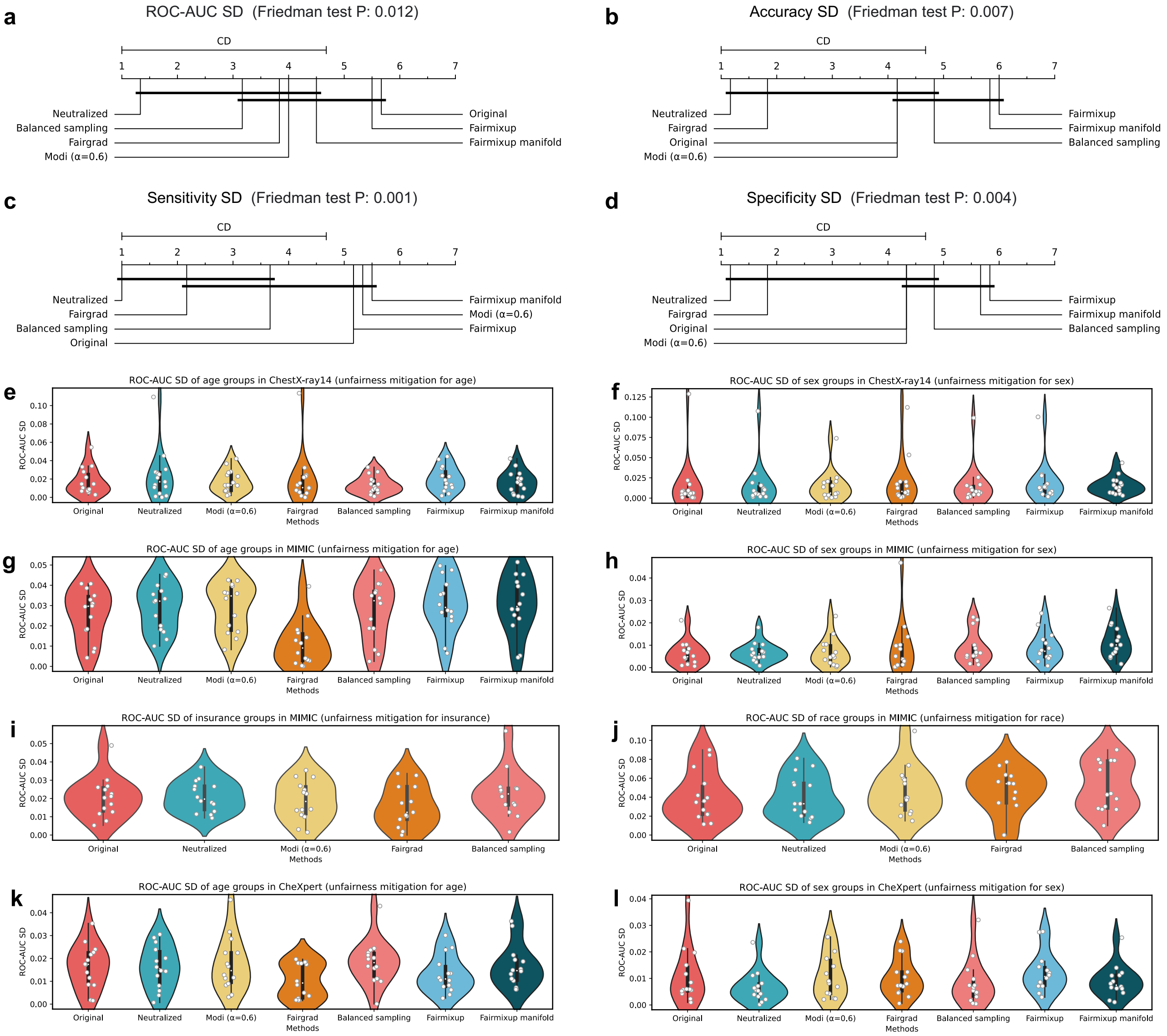

Fig. 6 | Unfairness comparison of various unfairness mitigation algorithms in disease diagnosis. Critical Difference (CD) diagrams for ROC-AUC SD (**a**), accuracy SD (**b**), sensitivity SD (**c**), and specificity SD (**d**). In each diagram, the Friedman test and the Nemenyi post-hoc test are performed across 6 {dataset, attribute} combinations: {ChestX-ray14, age}, {ChestX-ray14, sex}, {MIMIC-CXR, age}, {MIMIC-CXR, sex}, {CheXpert, age}, and {CheXpert, sex}. The CD value is 3.68. Violin plots of ROC-AUC SD for various DDMs in ChestX-ray14 (**e**, **f**), MIMIC-CXR (**g**–**j**), and CheXpert (**k**, **l**). The violin plot shows the distribution of ROC-AUC SDs across all findings (15 findings in ChestX-ray14, 14 findings in MIMIC-CXR, and 14 findings in CheXpert). The attributes corresponding to unfairness mitigation include age (**e**, **g**, **k**), sex (**f**, **h**, **l**), insurance (**i**), and race (**j**). In the violin plot, the central white dot represents the median, while the thick line inside the violin indicates the interquartile range. The whiskers represent the range of the data, excluding outliers. Source data are provided as a Source Data file.

neutralized-based DDMs on multiple metrics, such as ROC-AUC SD and worst-case ROC-AUC (Fig. 6e–l). However, the Fairmixup-based and Fairmixup-manifold-based DDMs generally exhibited worse performance across all 12 unfairness metrics and even performed inferiorly compared to the original-based DDM on certain unfairness metrics. For a more comprehensive unfairness evaluation, please refer to Supplementary Tables S18–20.

The Pearson correlation coefficient is utilized to quantify the correlation between two unfairness metrics, with Pearson's r values presented in Supplementary Table S21. As depicted in Supplementary Table S21, a negative correlation is observed between the worst-case metric and the other two types of unfairness metrics, with varying strengths observed. Conversely, there exists a strong positive correlation between the performance gap and the standard deviation (SD) across all four metrics: ROC-AUC (Pearson's r: 0.7195), accuracy (Pearson's r: 0.8794), sensitivity (Pearson's r: 0.9437), and specificity (Pearson's r: 0.7219).

Unfairness in multi-attributes neutralization

The population has multiple attributes, resulting in multiple sources of unfairness. This section aims to explore the impact of multi-attribute AttrNzrs on the performance and unfairness of DDMs. Four types of neutralized X-ray images were generated using the AttrNzr, where the neutralized attributes were (sex), (sex, age), (sex, age, race), and (sex, age, race, insurance), respectively. Subsequently, five DDMs are trained based on the original X-ray images and these four types of neutralized X-ray images individually. Figure 7 shows the ROC-AUCs and sensitivity SDs of these five DDMs on three datasets. It should be noted that the neutralized attribute and the evaluated attribute for unfairness may be different. For example, in Fig. 7k, the evaluated

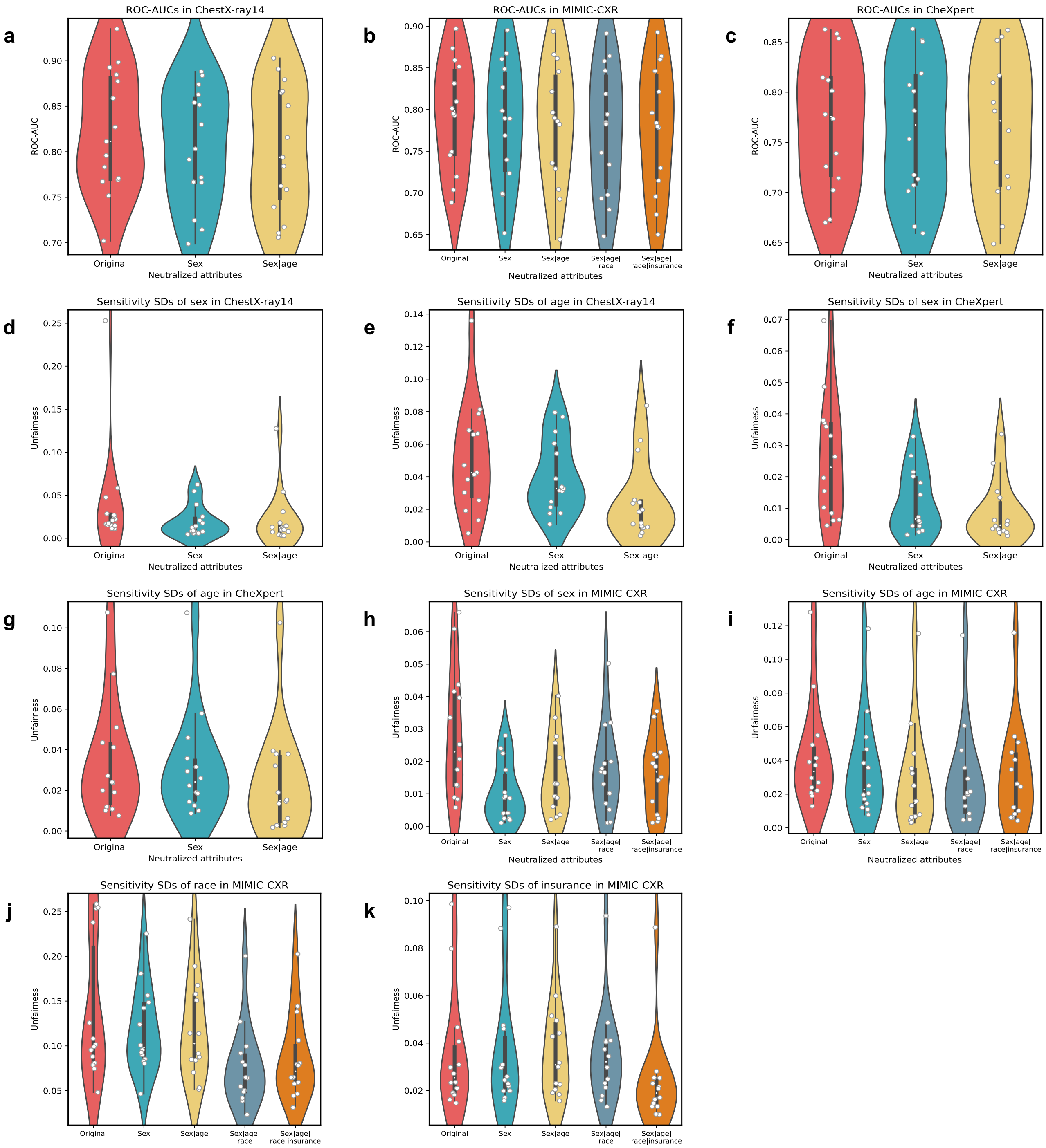

Fig. 7 | Violin plots of ROC-AUCs and sensitivity SDs of DDMs. The violin plot shows the distribution of ROC-AUCs or sensitivity SDs across all findings (15 findings in ChestX-ray14, 14 findings in MIMIC-CXR, and 14 findings in CheXpert). The distribution of ROC-AUCs in ChestX-ray14 (**a**), MIMIC-CXR (**b**), and CheXpert (**c**). The distribution of sensitivity SDs in ChestX-ray14 (**d**, **e**), MIMIC-CXR (**h**, **i**, **j**, **k**), and CheXpert (**f**, **g**). The attributes assessed for unfairness are as follows: sex (**d**, **f**, **h**), age (**e**, **g**, **i**), race (**j**), and insurance (**k**). There are multiple DDMs trained using either original X-ray images or neutralized X-ray images. The neutralized attributes include (sex), (sex, and age), (sex, age, and race), and (sex, age, race, and insurance). In the violin plot, the central white dot represents the median, while the thick line inside the violin indicates the interquartile range. The whiskers represent the range of the data, excluding outliers. Note: The neutralized attribute of the model training data and the attributes where unfairness is assessed may differ. For instance, in subfigure **k**, the assessed attribute is insurance, but only one model is trained on data with insurance neutralized. The ROC-AUCs and sensitivity SDs of each finding are shown in Supplementary Figs. S14 and S15. Source data are provided as a Source Data file.

attribute for unfairness is insurance, but insurance is neutralized only in X-ray images where sex, age, race, and insurance all are neutralized.

Illustrated in Fig. 7a–c, there exists a marginal reduction in the Macro-ROC-AUC of the DDM as the number of neutralized attributes increases. For instance, in the context of MIMIC-CXR (Fig. 7b), the Macro-ROC-AUC values for the five DDMs are as follows: 79.41% (original), 78.26% (neutralized sex), 78.21% (neutralized sex, and age), 77.44% (neutralized sex, age, and race), and 77.76% (neutralized sex,

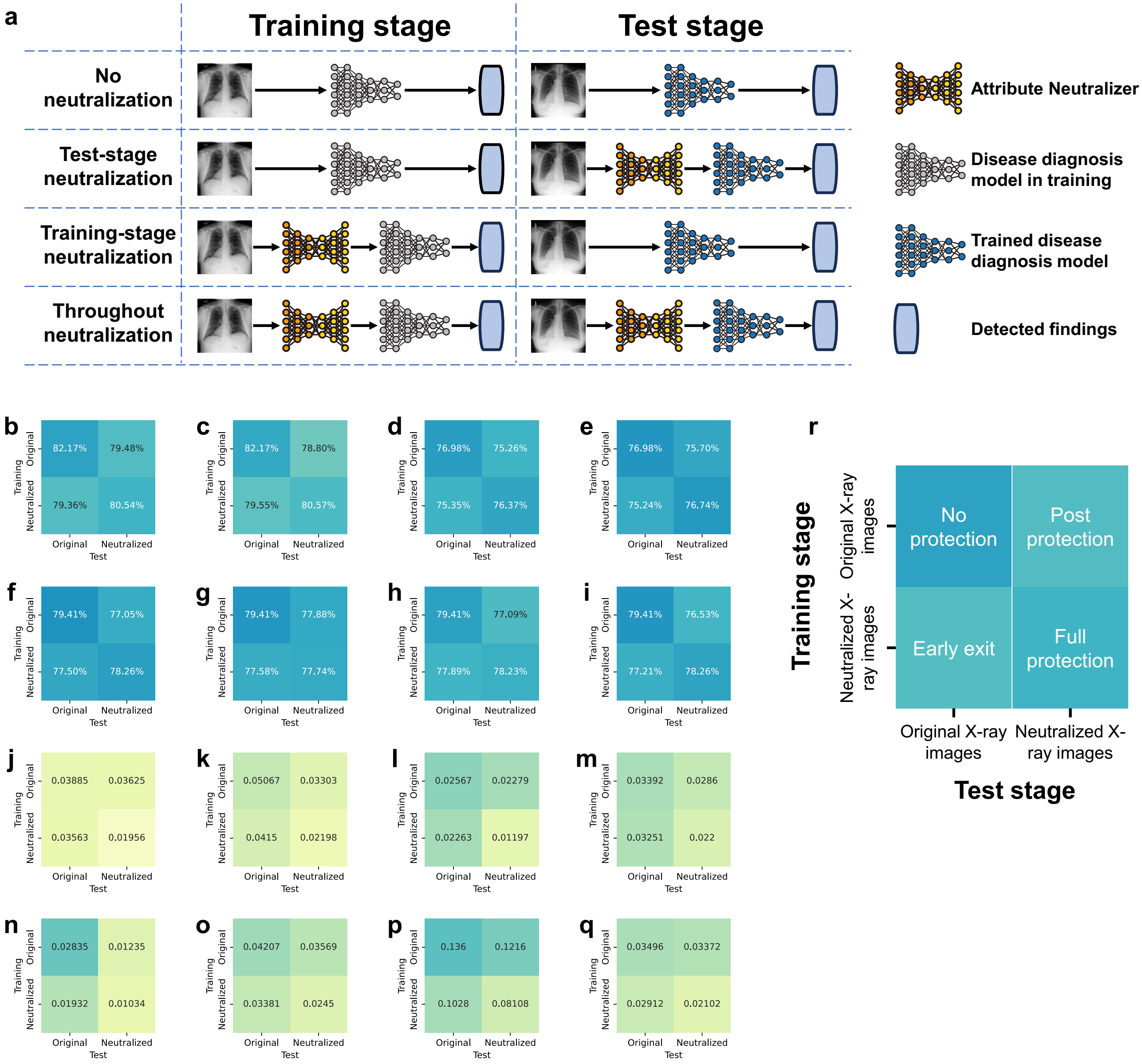

Fig. 8 | Macro-ROC-AUCs and sensitivity SDs of DDMs in four application paradigms. a Schematic diagram illustrating the four application paradigms of the AttrNzr. The four paradigms are as follows: no neutralization (no AttrNzr is used), test-stage neutralization (AttrNzr is used only during the test stage), training-stage neutralization (AttrNzr is used only during the training stage), throughout neutralization (AttrNzr is used during both the training and test stages). Macro-ROC-AUCs in ChestX-ray14 (**b**, **c**), CheXpert (**d**, **e**), and MIMIC-CXR (**f**–**i**). **d**–**k** Sensitivity SDs in ChestX-ray14 (**j**, **k**), CheXpert (**l**, **m**), and MIMIC-CXR (**n**–**q**). The neutralized attributes for each subfigure are as follows: sex (**b**, **d**, **f**, **j**, **l**, **n**), age (**c**, **e**, **g**, **k**, **m**, **o**), race (**h**, **p**), and insurance (**i**, **q**). **r** Detailed layout template for subgraphs **b**–**q**.

age, race, and insurance). This observation underscores that, like the single-attribute AttrNzr, the X-ray images generated by the multi-attribute AttrNzr can still retain sufficient disease-related information.

As depicted in Fig. 7d–k, when the neutralized attributes of the X-ray image encompass the attribute under evaluation, the reduction in unfairness related to the assessed attribute becomes evident. In Fig. 7e, the sensitivity SD for age decreased from 0.0507 (original) to 0.0249 (neutralized sex, and age). However, if the attribute under assessment is absent among the neutralized attributes of the X-ray image, the corresponding sensitivity SD displays minimal alteration. In Fig. 7j, the sensitivity SD for race shifted from 0.1360 (original) to 0.1184 (neutralized sex, and age); in Fig. 7k, the sensitivity SD for insurance changed from 0.0350 (original) to 0.0370 (neutralized sex, and age). This observation underscores the ability of the multi-attribute AttrNzr to alleviate unfairness in DDMs across numerous attributes. Simultaneously, this mitigation is highly targeted, with minimal influence on the unfairness of un-neutralized attributes.

Test-stage and training-stage neutralization paradigms

This section aims to explore two questions: 1) Can AttrNzrs offer post-protection of fairness for pre-existing models that were not originally designed to mitigate unfairness? 2) The parameters of the AttrNzr are approximately ten times larger than those of the DDM. To minimize the computational costs during the application, the AttrNzr is only introduced during the application stage. Can the early exit of the AttrNzr still effectively mitigate the unfairness of the model?

For the sake of explanation, four application paradigms of AttrNzr are defined: no neutralization (no AttrNzr is used), test-stage neutralization (AttrNzr is added only during the test stage of the pre-existing model), training-stage neutralization (AttrNzr is used only during the training stage of the new model), throughout neutralization (AttrNzr is used during both the training and test stages) (Fig. 8a). Macro-ROC-AUCs and average sensitivity SDs of DDMs in these four application paradigms are shown in Fig. 8b–q.

Illustrated in Fig. 8b–i, the average Macro-ROC-AUCs across the three datasets for the DDM in the four paradigms are as follows: 79.49% (no neutralization), 77.22% (test-stage neutralization), 77.46% (training-stage neutralization), and 78.34% (throughout neutralization). In the test-stage neutralization and training-stage neutralization paradigms, the reduction in the average Macro-ROC-AUC of the DDM might be attributed to the heterogeneity between the training data (training stage) and the test data (test stage). However, this decline is relatively modest, indicating that the performance of the DDM is not substantially affected in the test-stage neutralization and training-stage neutralization paradigms of the AttrNzr.

As shown in Fig. 8j–q, the sensitivity SDs across three datasets of the DDM in the four paradigms are 0.0488 (no neutralization), 0.0405 (test-stage neutralization), 0.0397 (training-stage neutralization), and 0.0266 (throughout neutralization). It can be seen that in the training-stage neutralization paradigms, the AttrNzr can still provide some fair protection effect. These results underscore how the introduction of the training-stage neutralization paradigms expands the potential applications of the AttrNzr.

Discussion

In this study, we proposed an attribute-neutral framework for mitigating unfairness in medical scenarios. Within this framework, we utilize AttrNzr to generate neutralized data. By employing neutral data in training DDMs, the improper correlation between disease information and sensitive attributes is effectively disrupted, thereby reducing the model's unfairness. Across three large public X-ray image datasets, AttrNzr demonstrates proficient reconstruction of X-ray images and accurate adjustment of attribute information intensity. Comparative analysis with other unfairness mitigation algorithms reveals that AttrNzr outperforms in multiple unfairness evaluation metrics. Furthermore, AttrNzr does not significantly reduce the diagnostic performance of the DDM across the entire population. Even when modifying multiple attributes, AttrNzr effectively mitigates model unfairness while preserving diagnostic performance. Lastly, AttrNzr proves effective in safeguarding model fairness in training-stage neutralization paradigms.

Among various DDMs, the Fairgrad-based DDM exhibits notable performance in terms of performance gap and performance SD (Fig. 6, and Supplementary Fig. S13). However, its disease diagnosis performance across the entire population is comparatively inferior (Fig. 4, and Supplementary Fig. S9). This phenomenon, referred to as leveling-down, deviates from the desired fairness standards in many practical scenarios[21,25,26]. Numerous in-processing unfairness mitigation algorithms introduce additional learning objectives to the loss function during training[11,12,14]. To some extent, this additional constraint conflicts with the model's primary objective, consequently leading to the prevalent occurrence of the leveling-down effect.

If a technique for mitigating unfairness in AI-enabled medical systems proves to be effective in real clinical scenarios, deployed AI-enabled medical systems may need to be scrapped and new systems developed based on the technology. In this case, the design, development, execution, testing, and deployment of system development will need to be redone. This will not only consume extra manpower and funds but also more medical resources. And these problems will be even more severe in underdeveloped regions. In this study, the test-stage neutralization paradigm of the AttrNzr may be one of the options to solve this problem. It can provide certain protection for the fairness of the system while retaining the original AI-enabled medical system.

In underdeveloped regions, medical resources are scarce and there is a lack of experienced doctors. Therefore, underdeveloped regions are the places where the advantages of AI-enabled medical systems can be best utilized. However, patients in underdeveloped regions are often an underrepresented population in AI training datasets. In the UK Biobank dataset, researchers found evidence of a "healthy volunteer" selection bias[27]. In the 23andMe genotype dataset of 2399 individuals, 2098 (87%) are European, while only 58 (2%) are Asian and 50 (2%) are African[28]. Therefore, AI-enabled medical systems deployed in underdeveloped regions may have more serious unfairness problems. On the other hand, computational resources are scarce in underdeveloped regions. Hence unfairness mitigation techniques that require heavy computational resources may be difficult to apply effectively. In this study, the training-stage neutralization paradigm of the AttrNzr does not require heavy computing resources. Since most patients in underdeveloped regions are underrepresented populations, the deployment of AttrNzrs can not only improve the fairness of the system but may also improve the overall diagnostic performance.

Our experimental results show that there is no significant correlation between certain unfairness evaluation metrics. For instance, the correlation between worst-case accuracy and accuracy gap (Pearson's r: −0.1767), as well as between worst-case ROC-AUC and ROC-AUC SD (Pearson's r: −0.20172), is weak. This underscores the incompatibility of various fairness definitions[29] at an experimental level. While performance gap, performance SD, and worst-case performance effectively gauge performance disparities among different groups, they may not directly capture the leveling-down effect. Given the inconsistency between performance and fairness, the complexity of individual attributes, and the groups of cross-attributes, adopting a comprehensive and diverse evaluation system is crucial for the development of effective unfairness mitigation algorithms.

The study has several limitations. Firstly, the current version of the AttrNzr can only modify discrete attributes. However, many attributes, such as age, income, etc., are continuous variables. To apply the AttrNzr to continuous attributes, they need to be discretized. Discretization introduces variance between groups of attributes. For instance, a patient aged 59 years and 364 days, and another aged 60 years and 1 day, would be grouped as "<60 y" and "≥60 y", respectively, despite the negligible real age difference. This forced distortion of the data distribution may prevent the AttrNzr from effectively learning the true change trend of continuous attributes. Secondly, the effectiveness of the AttrNzr should be validated on other types of imaging models and additional attributes. Thirdly, compared to worst-case performance and performance gap, the unfairness metric, performance SD, possesses less interpretability. Hence, future studies should consider this limitation when utilizing performance SD. While the current study demonstrated promising results on DDMs and specific attributes, it remains crucial to explore its performance across different medical imaging applications and various attribute domains.

Methods

Datasets

In this study, we include three large-scale public chest X-ray datasets, namely ChestX-ray14[15], MIMIC-CXR[16], and CheXpert[17]. The ChestX-ray14 dataset comprises 112,120 frontal-view chest X-ray images from 30,805 unique patients collected from 1992 to 2015 (Supplementary Table S1). The dataset includes 14 findings that are extracted from the associated radiological reports using natural language processing (Supplementary Table S2). The original size of the X-ray images is 1024 × 1024 pixels. The metadata includes information on the age and sex of each patient.

The MIMIC-CXR dataset contains 356,120 chest X-ray images collected from 62,115 patients at the Beth Israel Deaconess Medical Center in Boston, MA. The X-ray images in this dataset are acquired in one of three views: posteroanterior, anteroposterior, or lateral. To ensure dataset homogeneity, only posteroanterior and anteroposterior view X-ray images are included, resulting in the remaining

239,716 X-ray images from 61,941 patients (Supplementary Table S1). Each X-ray image in the MIMIC-CXR dataset is annotated with 13 findings extracted from the semi-structured radiology reports using a natural language processing tool (Supplementary Table S2). The metadata includes information on the age, sex, race, and insurance type of each patient.

The CheXpert dataset consists of 224,316 chest X-ray images from 65,240 patients who underwent radiographic examinations at Stanford Health Care in both inpatient and outpatient centers between October 2002 and July 2017. The dataset includes only frontal-view X-ray images, as lateral-view images are removed to ensure dataset homogeneity. This results in the remaining 191,229 frontal-view X-ray images from 64,734 patients (Supplementary Table S1). Each X-ray image in the CheXpert dataset is annotated for the presence of 13 findings (Supplementary Table S2). The age and sex of each patient are available in the metadata.

In all three datasets, the X-ray images are grayscale in either ".jpg" or ".png" format. To facilitate the learning of the deep learning model, all X-ray images are resized to the shape of 256×256 pixels and normalized to the range of [−1, 1] using min-max scaling. In the MIMIC-CXR and the CheXpert datasets, each finding can have one of four options: "positive", "negative", "not mentioned", or "uncertain". For simplicity, the last three options are combined into the negative label. All X-ray images in the three datasets can be annotated with one or more findings. If no finding is detected, the X-ray image is annotated as "No finding".

Regarding the patient attributes, the age groups are categorized as "<60 years" or "≥60 years"[30]. The sex attribute includes two groups: "male" or "female". In the MIMIC-CXR dataset, the "Unknown" category for race is removed, resulting in patients being grouped as "White", "Hispanic", "Black", "Asian", "American Indian", or "Other". Similarly, the "Unknown" category for insurance type is removed and patients are grouped as "Medicaid", "Medicare", or "Other". The amount and proportion of X-ray images under attributes and cross-attributes for the three datasets are shown in Supplementary Tables S1, S3–S5.

All three large-scale public chest X-ray datasets are divided into training datasets, validation datasets, and test datasets using an 8:1:1 ratio (Supplementary Table S6). To prevent label leakage, X-ray images from the same patient are not assigned to different subsets.

Attribute neutralizer

The AttrNzr is structured based on AttGAN[31], allowing for continuous adjustment of attribute intensity while preserving other image information. It consists of two main components: the generator and the discriminator. The generator employs a U-net structure to encode the original X-ray image as a latent representation and decodes the concatenation of the latent representation and the attribute vector into the modified X-ray image. The discriminator serves as a multi-task image classifier, distinguishing between the original and modified X-ray images while identifying the X-ray attribute. The AttrNzr's parameters are optimized through a loss function that combines attribute classification constraints, reconstruction loss, and adversarial loss (Supplementary Fig. S1).

G_{enc} and G_{dec} indicate the encoder and the decoder of the generator. C and D indicate the attribute classifier and the discriminator. Denoted by a the original attribute vector, b the modified attribute vector, $\hat{b}$ the identified attribute vector by C, Z the latent representation, x^a the original X-ray image with a, $x^{\hat{a}}$ the modified X-ray image with a, and $x^{\hat{b}}$ the modified X-ray image with b. a, b, and $\hat{b}$ contain n binary attributes, and can be expressed as $a = (a_1, \cdots, a_n)$, $b = (b_1, \cdots, b_n)$, and $\hat{b} = \left(\hat{b}_1, \cdots, \hat{b}_n\right)$, respectively.

In the AttrNzr, the image generated by the generator (encoder, and decoder) should meet three objectives: 1) $x^{\hat{a}}$ is the same as x^a; 2) the attribute of $x^{\hat{b}}$ is identified by C as b; and 3) $x^{\hat{b}}$ is identified by D as the real X-ray image. Therefore, the loss function of the generator L_{gen} is formulated as follows:

$$L_{gen} = \lambda_1 L_{rec} + \lambda_2 L_{cls_g} + L_{adv_g}, \tag{1}$$

where L_{rec}, L_{cls_g}, and L_{adv_g} indicate the reconstruction loss, the attribute classification constraint, and the adversarial loss, respectively. λ_1 and λ_2 are hyperparameters for balancing different losses. L_{rec} is measured by the sum of all the absolute differences between x^a and $x^{\hat{a}}$, and is formulated as follows:

$$L_{rec} = ||x^a - x^{\hat{a}}||_1. \tag{2}$$

L_{cls_g} is measured by the cross entropy between b and $\hat{b}$, and is formulated as follows:

$$L_{cls_g} = \sum_{i=1}^{n} -b_i log C_i\left(x^{\hat{b}}\right) - (1 - b_i) log\left(1 - C_i\left(x^{\hat{b}}\right)\right), \tag{3}$$

where $C_i\left(x^{\hat{b}}\right)$ indicates the predication of the i^{th} attribute. L_{adv_g} is formulated as follows:

$$L_{adv_g} = -D\left(x^{\hat{b}}\right), \tag{4}$$

In the AttrNzr, the discriminator/attribute-classifier should meet three objectives: 1) identify the attributes of x^a as a; 2) identify x^a as the real X-ray image; and 3) identify $x^{\hat{b}}$ as the fake X-ray image. Therefore, the loss function of the discriminator/attribute-classifier $L_{dis/cls}$ is formulated as follows:

$$L_{dis/cls} = \lambda_3 L_{cls_c} + L_{adv_d}, \tag{5}$$

where L_{cls_c} and L_{adv_d} indicate the attribute classification constraint, and the adversarial loss, respectively. λ_3 is the hyperparameter for balancing different losses. L_{cls_c} is measured by the cross entropy between a and the attribute vector produced by C, and is formulated as follows:

$$L_{cls_c} = \sum_{i=1}^{n} -a_i log C_i(x^a) - (1 - a_i) log(1 - C_i(x^a)). \tag{6}$$

L_{adv_d} is formulated as follows:

$$L_{adv_g} = -D(x^a) + D\left(x^{\hat{b}}\right). \tag{7}$$

The attribute vector comprises binary representations of attributes. For age and sex, "<60 years"/"≥60 years" and "female"/"male" are represented by 0/1. For multiclass attributes like race and insurance type, each subgroup is encoded using the one-hot encoding (Supplementary Fig. S2a). For example, the White is encoded as (1,0,0,0,0,0), and the Hispanic is encoded as (0,1,0,0,0,0). In the AttrNzr, the X-ray attribute is adjusted by modifying the attribute vector. The modification intensity α controls the degree of attribute modification. α ranges from 0 to 1, with 0 indicating no modification, 1 indicating negation of the attribute, and 0.5 indicating a neutral attribute (Supplementary Fig. S2b).

The high scalability of the attribute vector allows AttrNzr to modify not only a single attribute but also multiple attributes simultaneously. For the three chest X-ray datasets, single-attribute AttrNzrs and multi-attribute AttrNzrs are trained respectively (Supplementary Table S7).

To enhance the fundamental stability of the AttrNzr, several tips are implemented: 1) Gaussian noise with a mean of 0.1 is added to the X-ray image before inputting it to the discriminator; 2) 5% of fake/real labels are flipped during discriminator training; 3) label smoothing is applied to the attribute vector; 4) random horizontal flips are used to augment the X-ray image dataset; 5) a relatively large convolution kernel of size 6 × 6 is utilized; 6) the loss weights for the attribute classification constraint, reconstruction loss, adversarial loss, and gradient penalty are set to 10, 100, 1, and 10, respectively. Other training hyperparameters include a learning rate of 0.0001, a batch size of 64, and a training epoch of 300. The AttrNzr is trained on a Tesla V100 32GB GPU.

AI judge for attribute recognition

In this study, the judges identify the attributes of X-ray images generated by our AttrNzr. The first judge is an AI model that has been fully trained on original X-ray images to classify attribute types. The AI judge is used to identify the attributes of X-ray images that have been modified with different intensities. The modification intensity α is set to 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. To facilitate the evaluation of the performance of the AI judge, only AI judges for binary attributes (age, and sex) are trained in this study.

After considering the performance of various deep learning models in disease diagnosis (which will be mentioned in the Disease Diagnosis Model section), ConvNet[32] is selected to build the AI judge. The AI judge is designed with 2 output nodes, corresponding to "<60 years"/"≥60 years" or "female"/"male". All parameters of the AI judge are initialized using ConvNet's pre-training on the ImageNet dataset. Data augmentation techniques[33], including random horizontal flip, random rotation, Gaussian blur, and random affine, are applied to expand the dataset. Other hyperparameters include a learning rate of 0.0005, a batch size of 120, and a training epoch of 100. The AI judge is also trained on a Tesla V100 32GB GPU. After the AI judge is fully trained, the Gradient-weighted Class Activation Map is involved to find the activated region of the modified X-ray image.

Human judge for attribute recognition

The second attribute recognition involves human judges identifying the attributes of X-ray images generated by our AttrNzr. Five junior physicians are invited from the Thoracic Surgery Department of Guangdong Provincial People's Hospital to act as human judges for the attribute recognition. Due to variations in race and insurance between the regions where the large-scale public chest X-ray datasets are acquired and the working regions of the 5 human judges, the attribute recognition focused only on the attributes of age and sex.

For each attribute, 5 groups of X-ray images are randomly selected from the ChestX-ray14 dataset. Each group contains 40 X-ray images that are modified by the AttrNzr using different modification intensities. To reduce the workload of the human judges, the modification intensity α is limited to five values: 0.0, 0.3, 0.5, 0.7, and 1.0.

Even with different modification intensities, the same group of X-ray images still exhibits relative similarity. To prevent the identification decisions of the human judges from being influenced by modified X-ray images of the same group but with different modification intensities, each judge is not allowed to repeatedly identify the same group, regardless of the modification intensity. The assignment schedule for the five human judges is presented in Supplementary Fig. S3.

Disease diagnosis model

After comparing the disease diagnosis performance of various deep learning networks on the three large-scale public chest X-ray datasets (Supplementary Table S8), ConvNet is selected as the DDM for this study. In these datasets, the "No finding" label and other finding labels are mutually exclusive, but the other finding labels themselves are not mutually exclusive. To simplify the disease diagnosis task, it is treated as a multi-label recognition task. In the DDM, the number of output nodes is equal to the number of finding labels, including the "No finding" label. The activation function of the last layer is sigmoid, and the loss function is binary cross-entropy, which calculates the loss between the target and the output probabilities. Taking into account the imbalance of findings in the dataset, we assign weights to the losses of the findings based on the number of X-ray images associated with each finding. The initialization, data augmentation, and hyperparameter settings remain consistent with those of the AI Judge.

The instability of deep learning poses uncertainty in the evaluation of DDMs. To ensure reliable evaluation results, we conduct additional training for 20 epochs after the DDM has converged on the validation dataset. At the end of each training epoch, we save the output of the DDM. Finally, the DDM is evaluated based on the outputs obtained from these 20 epochs.

Alternative unfairness mitigation algorithms

Three alternative algorithms for mitigating unfairness in AI-enabled medical systems are introduced in this study: the Fairmixup[12], the Fairgrad[11], and the Balanced sampling[18]. The first two algorithms require integration into the DDM, while the third is solely applied to the dataset.

In the Fairmixup, mixup is employed to generate interpolated samples between different groups[12]. These interpolated samples introduce a smoothness regularization constraint that is incorporated into the loss function of AI models to mitigate unfairness. Mixup can be implemented at both the image and feature levels, referred to as Fairmixup and Fairmixup manifold, respectively. Interpolated samples are derived from blending two samples, thus, Fairmixup is effective in addressing unfairness associated with binary attributes such as age and sex. The implementation of Fairmixup is based on the official algorithm source code (https://github.com/chingyaoc/fair-mixup), with the regularization constraint weight in the loss function set to 0.05.

The Fairgrad ensures fairness by assigning lower weights to examples from advantaged groups compared to those from disadvantaged groups[11]. This method is applicable only to binary classification tasks. Consequently, in our investigation, the multi-label recognition task is segmented into multiple binary classification tasks (15, 14, and 14 binary classification tasks in the ChestX-ray14, MIMIC-CXR, and CheXpert datasets respectively). The Fairgrad's implementation is based on the official PyPI package (https://pypi.org/project/fairgrad/), and unfairness in the loss function is assessed using equalized odds.

Balanced sampling combats unfairness by constructing group-balanced data, wherein the sample size of majority groups is randomly down-sampled to match that of the minority group while preserving proportional distributions among various findings. Details regarding the sample size of the minority group are available in Supplementary Table S1.

For each alternative unfairness mitigation algorithm, model framework, data augmentation, learning rate, number of training epochs, and other configurations remain consistent with the baseline DDM.

Performance evaluation metrics

The SSIM[34] is utilized to evaluate the similarity between two X-ray images. SSIM is calculated on various windows of an image. The measure between two windows x and y, with a size of $N \times N$, is given by the formula:

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)}. \tag{8}$$

Here, μ_x and μ_y represent the mean pixel values of x and y, respectively. σ_x^2 and σ_y^2 denote the variances of x and y, while σ_{xy} represents the cross-correlation between x and y. The variables c_1 and c_2 are included to stabilize the division when the denominator is weak. The size of the window is set to 100 × 100 in our study.

In attribute recognition, the performance of the AI judge in identifying the original attributes of the modified X-ray image is evaluated using accuracy, sensitivity, specificity, and F1 score. Additionally, the area under the receiver operating characteristic curve (AUC-ROC) is calculated to provide further evaluation of the AI judge. For the human judges, only accuracy is used to assess their performance in identifying the original attributes of the modified X-ray image.

To address the instability of the DDM, the outputs obtained from 20 epochs after convergence are averaged to obtain a stable output. In assessing the performance of the DDM for each finding, ROC curves and precision-recall (PR) curves are generated, and corresponding AUC values are computed. Additionally, accuracy, sensitivity, specificity, precision, and F1-score are calculated for evaluation purposes. Macro-averaging of these metrics across all findings is performed to assess the overall performance of the DDM.

Unfairness evaluation metrics

Unfairness is assessed by examining the performance of various subgroups[20]. In our study, ROC-AUC serves as the primary metric for evaluating model performance. To assess unfairness related to non-binary attributes, we employ two evaluation metrics: (1) Group Fairness, which measures the gap in ROC-AUC between subgroups with the highest and lowest AUC values[20], and (2) Max-Min Fairness, which evaluates the AUC of the subgroup with the poorest performance[20,21]. Furthermore, we report values for other performance metrics such as accuracy, sensitivity, and specificity.

Neither the Worst-case ROC-AUC nor the ROC-AUC Gap can reflect the performance differences among all subgroups. The standard deviation (SD) can measure the mutual difference among multiple variables. Therefore, we introduce the standard deviation of performance[13,22] as the third evaluation metric of unfairness. The performance SD can be computed using the following formula:

$$UI = \sqrt{\frac{\sum_{i=1}^{M}\left(met_i - \overline{met}\right)^2}{M}}. \tag{9}$$

Here, M represents the number of groups in the attribute. met_i denotes the performance metric value of the ith group, and $\overline{met}$ represents the average performance across all groups.

The performance SD quantifies the variation or dispersion of performance among different groups within the attribute. A low performance SD suggests that the performance of each group is closer to the average performance across all groups, indicating less unfairness. Conversely, a high performance-SD suggests that the performance of each group is spread out over a wider range, indicating greater unfairness. In our study, the calculation of the performance SD is also based on the stable output of the DDM.

Statistical analysis

SSIM is utilized to assess the similarity between the modified X-ray image and the original X-ray image. Subsequently, the Pearson correlation coefficient is employed to measure the correlation between the similarity and modification intensity. Additionally, the Pearson correlation coefficient is also used to evaluate the correlation between the judges' identification performance and the modification intensity. The evaluation metrics of the DDM are calculated at a 95% confidence interval using non-parametric bootstrapping with 1000 iterations. Delong's test is employed to test the statistical significance of the difference between two ROC curves. The confidence interval for the difference between the areas under the PR curves is computed using the bias-corrected and accelerated bootstrap method. If the 95% confidence interval does not encompass 0, it signifies a significant difference between the two areas ($P < 0.05$). The comparison of ROC curve and PR curve was performed by MedCalc.

For a comprehensive comparison of the relative performance and unfairness mitigation among different algorithms, we employ the Friedman test[35] followed by the Nemenyi post-hoc test[20]. Initially, relative ranks are computed for each algorithm within each dataset and attribute independently. Subsequently, if the Friedman test reveals statistical significance, the average ranks are utilized for the Nemenyi test. A significance threshold of $P < 0.05$ is adopted. The outcomes of these tests are presented through Critical Difference (CD) diagrams[36]. In these diagrams, methods connected by a horizontal line belong to the same group, indicating nonsignificant differences based on the p-value, while methods in distinct groups (not connected by the same line) exhibit statistically significant disparities. The Fairmixup and Fairmixup manifold techniques are unsuitable for non-binary attributes. Consequently, the Friedman test and the Nemenyi post-hoc test are performed across 6 {dataset, attribute} combinations: {ChestX-ray14, age}, {ChestX-ray14, sex}, {MIMIC-CXR, age}, {MIMIC-CXR, sex}, {CheXpert, age}, and {CheXpert, sex}.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

In this study, we incorporate three extensive public chest X-ray datasets: ChestX-ray14, MIMIC-CXR, and CheXpert. The ChestX-ray14 dataset is supported by the Intramural Research Program of the NIH Clinical Center and accessible at https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345. The MIMIC-CXR dataset is accessible at https://physionet.org/content/mimic-cxr-jpg/2.0.0/. To access the data files of the MIMIC-CXR dataset, one must first be a credentialed user of PhysioNet. Subsequently, the completion of mandatory training, such as CITI Data or Specimens Only Research, is required. Lastly, the data use agreement for the project should be signed. Notably, the chest X-ray images within the MIMIC-CXR dataset have been preprocessed into compressed JPG format. The initial chest X-rays in DICOM format can be retrieved at https://physionet.org/content/mimic-cxr/2.0.0/. The CheXpert dataset can be accessed through https://stanfordmlgroup.github.io/competitions/chexpert/. Source data are provided with this paper.

Code availability

The code for the AttrNzr, AI judge, and DDM can be accessed via the following link: https://zenodo.org/records/13254099[37]. Comprehensive instructions are provided to facilitate the replication of our work. For code testing purposes, we have made available reduced-scale data files. Moreover, all model hyperparameters are encompassed within the respective scripts. Notably, a segment of the AttrNzr code draws inspiration from the work of Elvis Yu-Jing Lin (https://github.com/elvisyjlin/AttGAN-PyTorch); the implementation of Fairmixup is derived from the official algorithm source code (https://github.com/chingyaoc/fair-mixup); while the implementation of Fairgrad is based on the official PyPI package (https://pypi.org/project/fairgrad/).

References

1. Kermany, D. S. et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **172**, 1122–1131.e9 (2018).
2. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
3. Li, D. et al. A proposed artificial intelligence workflow to address application challenges leveraged on algorithm uncertainty. *iScience* **25**, 103961 (2022).
4. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
5. Brown, A. et al. Detecting shortcut learning for fair medical AI using shortcut testing. *Nat. Commun.* **14**, 4314 (2023).

6. Char, D. S., Shah, N. H. & Magnus, D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).
7. Chen, I. Y., Joshi, S. & Ghassemi, M. Treating health disparities with artificial intelligence. *Nat. Med* **26**, 16–17 (2020).
8. Gichoya, J. W. et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* **4**, e406–e414 (2022).
9. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med* **27**, 2176–2182 (2021).
10. Pagano, T. P. et al. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data Cogn. Comput.* **7**, 15 (2023).
11. Maheshwari, G. & Perrot, M. *FairGrad: Fairness Aware Gradient Descent* (arXiv preprint, 2022).
12. Chuang, C.-Y. & Mroueh, Y. *Fair Mixup: Fairness via Interpolation* (arXiv preprint, 2021).
13. Puyol-Antón, E. et al. Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12903 (LNCS, 2021).
14. Dash, S., Balasubramanian, V. N. & Sharma, A. Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals. In *Proceedings – 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022* (IEEE, 2022). https://doi.org/10.1109/WACV51458.2022.00393.
15. Wang, X. et al. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 vols 2017-January* (IEEE, 2017).
16. Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
17. Irvin, J. et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence* (EAAI, 2019). https://doi.org/10.1609/aaai.v33i01.3301590.
18. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl Acad. Sci.* **117**, 12592–12594 (2020).
19. Zhang, H. et al. Improving the Fairness of Chest X-ray Classifiers. In *Proceedings of Machine Learning Research* **174**, 204–233 (2022).
20. Zong, Y., Yang, Y. & Hospedales, T. *MEDFAIR: Benchmarking Fairness for Medical Imaging* (arXiv preprint, 2022).
21. Martinez, N., Bertran, M. & Sapiro, G. Minimax pareto fairness: A multi objective perspective. In *37th International Conference on Machine Learning, ICML 2020* vols PartF168147-9 (ICLM, 2020).
22. Wang, M. & Deng, W. Mitigating bias in face recognition using skewness-aware reinforcement learning. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 2020). https://doi.org/10.1109/CVPR42600.2020.00934.
23. Das, A., Anjum, S. & Gurari, D. Dataset bias: A case study for visual question answering. *Proc. Assoc. Inf. Sci. Technol.* **56**, 58–67 (2019).
24. Reddy, C. et al. Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics (NeurIPS, 2021).
25. Maheshwari, G., Bellet, A., Denis, P. & Keller, M. *Fair Without Leveling Down: A New Intersectional Fairness Definition* (2023).
26. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* **13**, https://doi.org/10.1038/s41467-022-32186-3 (2022).
27. Fry, A. et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
28. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **54**, https://doi.org/10.1145/3457607 (2021).
29. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Leibniz International Proceedings in Informatics,* 67 (LIPIcs, 2017).
30. Chaudhari, S. J. Methodology for Gender Identification, Classification and Recognition of Human Age. *Int. J. Comput. Appl.* **NCAC2015**, 5–10 (2015).
31. He, Z., Zuo, W., Kan, M., Shan, S. & Chen, X. AttGAN: Facial Attribute Editing by only Changing What You Want. *IEEE Trans. Image Process.* **28**, 5464–5478 (2019).
32. Liu, Z. et al. A ConvNet for the 2020s. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 2022).
33. Hu, L., Liang, H. & Lu, L. Splicing learning: A novel few-shot learning approach. *Inf. Sci.* **552**, 17–28 (2021).
34. Wang, Z., Simoncelli, E. P. & Bovik, A. C. Multiscale structural similarity for image quality assessment. in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers,* 1398–1402 (IEEE, 2003). https://doi.org/10.1109/ACSSC.2003.1292216.
35. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **32**, 675–701 (1937).
36. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).
37. Hu, L. et al, Enhancing Fairness in AI-Enabled Medical Systems with the Attribute Neutral Framework, Attribute-Neutralizer-for-medical-AI-system. *Zenodo*, https://doi.org/10.5281/zenodo.13254099 (2024).

Acknowledgements

The authors express gratitude to Zhenliang He, one of the creators of AttGAN, for his insightful discussions. The authors also acknowledge the thoracic surgeons of Guangdong Provincial People's Hospital for their support in conducting the attribute recognition. Additionally, the authors extend their appreciation to the high-performance cluster (HPC) at Guangdong Provincial People's Hospital, and the Bio-Med Informatics Research Center at Xinqiao Hospital for their valuable support in numerical computations. This study's financial support was derived from various sources, including the National Key Research and Development Program of China (2019YFB1404803 to Huiying Liang), the General Program of the National Natural Science Foundation of China (62076076 to Huiying Liang), the Excellent Young Scientists Fund (82122036 to Huiying Liang), the Tianyuan Fund for Mathematics (12326612 to Huiying Liang), Young Scientists Fund (81903331 to Chao Xiong), the Knowledge Innovation Specialized Project of Wuhan Science and Technology Bureau (2023020201010198 to Chao Xiong), the Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (2022B1212010011 to Huiying Liang), the Basic and Applied Basic Research Foundation of Guangdong Province (2022A1515110722, and 2024A1515011750 to Lianting Hu), and the Guangdong Provincial Medical Science and Technology Research Fund Project (A2023011 to Lianting Hu).

Author contributions

Study conception and design: Lianting Hu, Dantong Li, and Huiying Liang; Data collection and analysis: Lianting Hu, Shuai Huang, and Xiaoting Peng; Interpretation of results: Lianting Hu, Dantong Li, Huiying

Liang, Huazhang Liu, Xueli Zhang, Xiaohe Bai; Manuscript preparation: Lianting Hu, Dantong Li, Huazhang Liu, Huan Yang, Lingcong Kong, Jiajie Tang, Peixin Lu; Manuscript review: Huiying Liang, Xuanhui Chen, Yunfei Gao, Chao Xiong; All authors reviewed the results and approved the final version of the manuscript. Lianting Hu, Dantong Li contributed equally to this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52930-1.

Correspondence and requests for materials should be addressed to Chao Xiong or Huiying Liang.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]The Data Center, Wuhan Children's Hospital (Wuhan Maternal and Child Healthcare Hospital), Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430016 Hubei, China. [2]Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Guangzhou 510080 Guangdong, China. [3]Medical Big Data Center, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080 Guangdong, China. [4]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Guangzhou 510080 Guangdong, China. [5]Medical Research Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080 Guangdong, China. [6]Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080 Guangdong, China. [7]School of Physical Sciences, University of California San Diego, La Jolla, San Diego, CA 92093, USA. [8]Clinical Medical Research Center, Xinqiao Hospital, Army Medical University, Chongqing 400037, China. [9]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA. [10]These authors contributed equally: Lianting Hu, Dantong Li. ✉e-mail: xcpopo11@yeah.net; lianghuiying@hotmail.com