Unbiasing Fairness Evaluation of Radiology AI Model

Yuxuan Liang[a], Hanqing Chao[a], Jiajin Zhang[a], Ge Wang[a], Pingkun Yan[a]

[a]*Biomedical Imaging Center, Center for Biotechnology and Interdisciplinary Studies, Department of Biomedical Engineering, Rensselaer Polytechnic Institute, 110 8th St, Troy, 12180, New York, United States*

Abstract

Fairness of artificial intelligence and machine learning models, often caused by imbalanced datasets, has long been a concern. While many efforts aim to minimize model bias, this study suggests that traditional fairness evaluation methods may be biased, highlighting the need for a proper evaluation scheme with multiple evaluation metrics due to varying results under different criteria. Moreover, the limited data size of minority groups introduces significant data uncertainty, which can undermine the judgement of fairness. This paper introduces an innovative evaluation approach that estimates data uncertainty in minority groups through bootstrapping from majority groups for a more objective statistical assessment. Extensive experiments reveal that traditional evaluation methods might have drawn inaccurate conclusions about model fairness. The proposed method delivers an unbiased fairness assessment by adeptly addressing the inherent complications of model evaluation on imbalanced datasets. The results show that such comprehensive evaluation can provide more confidence when adopting those models.

Keywords: Fairness, Deep learning, Evaluation metrics, Data uncertainty, Medical imaging.

1. Introduction

In recent years, the fairness of machine learning models has become a critical concern [1], especially in medical imaging tasks such as disease diagnosis [2, 3], organ segmentation [4], and image registration [5]. A commonly used criterion for assessing fairness in these areas is "equalized odds," which requires consistent model performance across different demographic cohorts, (e.g., specific races and genders) [6]. The most frequent cause for

the deviation from equalized odds is the imbalance in datasets, where some demographic groups are significantly underrepresented [7, 8]. This discrepancy is exacerbated in medical data due to barriers in access and disparities in healthcare standards, leading to larger discrepancies compared to general datasets [9]. As a result, models trained on these skewed datasets typically show higher accuracy for majority groups but underperform for minority groups. This disparity can lead to higher rates of misdiagnosis in underrepresented groups, causing harm not only to individuals but also to society as a whole [10].

Most research on model fairness to date has focused on mitigating the performance gap between the majority and the minority groups. However, this paper underscores the necessity of conducting a thorough fairness evaluation before implementing mitigation strategies. Traditional equalized odds based methods emphasize model accuracy comparison among different subgroups. Yet, in the medical imaging domain, relying solely on accuracy can be misleading, as other crucial metrics like sensitivity, specificity, and the F1 score also play significant roles. Experiments in this study demonstrate that under different evaluation metrics, the judgment of model fairness might differ or even become contradictory. For instance, one subgroup with higher accuracy than another might exhibit a lower F1 score. Such discrepancies are more likely in datasets with imbalanced labels, a common issue in medical imaging [11]. Considering this fact, and that the Receiver Operating Characteristic (ROC) curve comprehensively reflects model performance, this paper adopts the ROC curve as its evaluation metric. Our study further demonstrates that extreme dataset imbalance additionally hampers the accurate evaluation of model fairness, misleadingly suggesting fairness when it is not. To address this critical gap, an evaluation methodology is designed to offer a more holistic assessment of model fairness.

Traditional approaches to assessing model fairness involve separate evaluations on different demographic groups to identify any disparities in performance. However, the reliability of such assessments can be compromised by the limited size of test sets for minority groups. Although cross-validation [12] and bootstrapping [13] are commonly employed strategies for small datasets, previous studies have highlighted the uncertainty associated with these methods in biomedical contexts, where sample sizes are only in the hundreds [14]. This uncertainty introduces a significant margin of error in the performance evaluation of minority groups, rendering it challenging to ascertain whether observed performance gaps come from model bias or the inherent uncertainty

of the testing process. To circumvent this issue, our study introduces an approach that leverage bootstrapping of data from the majority group. The architecture our proposed evaluation method on ROC curves is shown in Figure 1. When the sample sizes of two groups are similar, conventional bootstrapping method can be adapted to fairness evaluation, in which both groups are bootstrapped the same number of samples as their original data sizes. However, in scenarios where there is a substantial disparity in the sizes of the majority and minority groups, modifications to the traditional bootstrapping approach are necessary to accommodate the significant differences in data sizes. The core concept involves simulating a distribution of test results by bootstrapping the majority group data to match the size of the minority group data. This enables statistical testing to determine whether the performance of the minority groups deviates significantly from that of the majority group simulations. A comparative analysis of ROC curves from both groups is undertaken [15], which involves testing both hypotheses of the majority group outperforming the minority group and vice versa. A model is considered to exhibit fairness issues if there is a significant imbalance in the occurrence of small p-values favoring one hypothesis over the other.

2. Materials and Method

2.1. Problem Definition

In this section, we focus on the situation when a large difference in the number of data samples exists between different groups. Given an imbalanced dataset $S = \{\{x_i, y_i, z_i\}, i \in 1, ..., T\}$, where x_i is a data sample, y_i is the class label, and $z_i \in \{0, 1\}$ is the demographic group. Suppose $\{z_i = 0\}$ represents the majority group, which contains M samples in total, and $\{z_i = 1\}$ represents the minority group with N samples. A classification model $f_\theta(\cdot)$ trained on this dataset may exhibit performance discrepancies between subgroups in the absence of bias mitigation strategies. The concept of equalized odds is mathematically defined as

$$P(f_\theta(x_i) = y_i | z_i = 0) = P(f_\theta(x_i) = y_i | z_i = 1). \tag{1}$$

A model is deemed unfair if there exists a substantial difference in accuracy across demographic groups, quantified as

$$|P(f_\theta(x_i) = y_i | z_i = 0) - P(f_\theta(x_i) = y_i | z_i = 1)| \geq \epsilon, \tag{2}$$

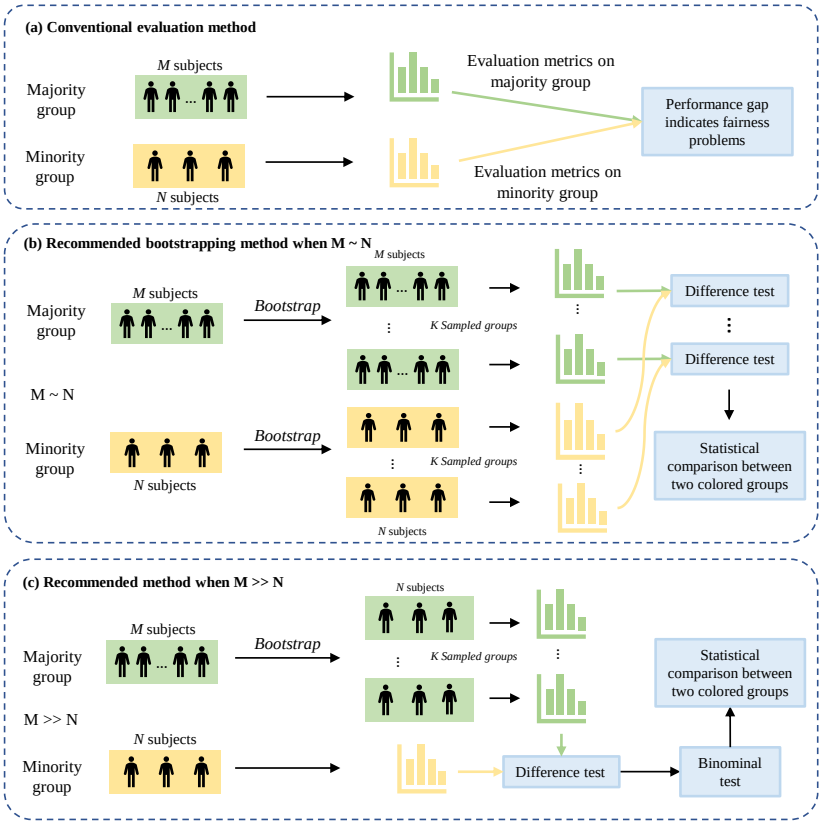


Figure 1: Comparison of (a) conventional evaluation method, (b) recommended bootstrapping method when M and N have similar values, and (c) proposed evaluation method when $M \gg N$.

where ϵ is a predefined threshold. In medical imaging domain, however, accuracy should not be the only consideration. The definition of equalized odds is expended into

$$|F(f_\theta(x_i, z_i = 0)) - F(f_\theta(x_i, z_i = 1))| \geq \epsilon, \tag{3}$$

where $F(f_\theta(x_i))$ can be any evaluation metrics. Nonetheless, the evaluation of performance in the minority group is often marred by significant uncertainty due to small sample sizes:

$$\tilde{F}(f_\theta(x_i, z_i = 1)) = F(f_\theta(x_i, z_i = 1)) \pm \delta, \tag{4}$$

where δ represents the uncertainty. Consequently, the observed performance gap, if any, becomes

$$|F(f_\theta(x_i), z_i = 0)) - \tilde{F}(f_\theta(x_i), z_i = 1))| \geq \epsilon. \tag{5}$$

Given that the relationship between uncertainty δ and fairness threshold ϵ is not clearly established, determining the fairness of the model remains challenging. The subsequent sections will explore methods to approximate δ through bootstrapping techniques applied to the majority group, thereby facilitating a more accurate evaluation of the model's fairness. A hypothesis is made that when the sample size is the same, a fair model should have the same uncertainty on both the majority and minority groups. Importantly, this hypothesis does not necessitate identical feature distributions or characteristics between the two groups. Instead, it demands that the model demonstrates similar performance across these groups when excluding the influence of sample size.

2.2. Fairness Evaluation

We discuss the evaluation in two different scenarios. The first situation is when the evaluation metric is a scalar value and the second one involves the compound metrics like Receiver Operating Characteristic (ROC) curves. When the evaluation metric is a scalar value F_j, such as accuracy or sensitivity, consider a testing dataset where the majority group comprises M samples and the minority group comprises N with $M \gg N$. The process starts with bootstrapping N samples from the majority group and computing the performance metric

$$\hat{F}_j(f_\theta(x_i, z_i = 1)), j \in [1, \ldots, k] \tag{6}$$

for each iteration. This process is repeated k times to generate a series of performance values. It is reasonable to assume that the distribution of these bootstrapped performance values approximates a normal distribution. Consequently, a p-value can be calculated using a t-test for assessing the statistical significance of the difference. As k increases, the test approximates a z-test due to the large sample size. A small p-value indicates a statistically significant deviation of the minor group's performance from the bootstrapped performance distribution of the majority group, suggesting potential fairness issues within the model.

For evaluation metrics involving ROC curves, the approach differs from that of single-value metrics due to the nature of ROC curves, which is represented a series of points rather than a single value. In this scenario, the bootstrapping methodology remains consistent with that used for single-value metrics. That is bootstrapping the same number of cases as the minority from the majority. However, for each set of bootstrapped samples, an ROC curve is generated. Subsequently, an ROC z-test is conducted to compare the ROC curve of the minority group against each bootstrapped ROC curve from the majority group. This comparison is framed within two one-sided hypotheses: the majority group outperforms the minority group and the reverse scenario. As a result, for each bootstrapped sample, two p-values, p_1 and p_2 corresponding to two hypothesises are calculated. This procedure is repeated across all bootstrapped samples. If the frequency of either sufficiently small p_1 or p_2 values is significantly different, it indicates a potential fairness issue within the model. For example, if the frequency of p_1 is much higher than p_2, it indicates on more bootstrapped samples of the majority group, the model out performs the minority group. A Binomial test is then performed to judge whether there is significant difference between the two p-values [16, 17].

2.3. Dataset and Preprocessing

The National Lung Screening Trail (NLST) dataset contains 10,395 subjects who underwent lung cancer screening low-dose helical computed tomography (LDCT), which has also been used to detect cardiovascular disease (CVD) [18]. Each subject was labeled as either CVD-positive or CVD-negative to conduct CVD screening. The demographic breakdown of the dataset includes 9,704 White subjects, 309 Black or African American subjects, 178 Asian subjects, and a nominal count of subjects from other racial backgrounds. Within the context of this research, White subjects are designated as the majority group, while Black or African American subjects are identified as the minority group. Each subject underwent several CT exams, and in total there are 41871 CT exams, 40,681 from White and 1,189 from Black or African American. A state-of-art segmentation model is applied to segment heart region from the whole 3D CT volumes, and the results are used to train deep learning models to predict CVD risks. The dataset categorizes the CT exam outcomes into two labels: '0' indicating a normal result and '1' signifying the presence of CVD. For the White group, 28412 exams are used

for training, and 12641 for test. For Black or African American group, 818 exams are used for training, 371 for test.

2.4. Implementation Details

Two deep neural network models were developed to implement the evaluation strategies proposed in this study. The first model is ResNet18 3D, which retains the original structure of ResNet18 but modifies the convolutional kernels to process 3D data[19, 20]. The second model, originating from the work of Chao *et al.* [18], employs an architecture of Tri-2D network consisting of three 2D ResNet branches. This model decomposes the original 3D CT volumes into 2D slices across three orthogonal planes - sagittal, coronal, and axial - and processes these slices separately in each ResNet18 branch. Additionally, an attention mechanism is incorporated within each branch to enhance feature extraction capabilities. The features extracted from the three branches are subsequently merged to produce the model's final output. Comparative analyses have shown that this multi-branch model with attention mechanisms surpasses the performance of the 3D ResNet18 model on the NLST dataset.Both models underwent training for 20 epochs with a batch size of 32.

3. Results

3.1. Evaluation with conventional method

Conventional evaluation outcomes were obtained under multiple evaluation metrics. Subsequent difference testing between the two subgroups highlighted in Table 1 reveals the average bootstrapping outcomes alongside the p-values from the difference assessments across several evaluation metrics. For both evaluated models, the White subgroup demonstrated significantly higher accuracy compared to the Black or African American subgroup. Conversely, metrics such as Precision, Recall, and F1 score favored the latter, presenting an inverse relationship. Table 2 delineates the confusion matrices for both the White and Black or African American subgroups concerning the ResNet 3*2D model specifically. These matrices elucidate a predominance of negative classes over positive ones within both groups, with the Black or African American subgroup exhibiting a notably low incidence of false negatives. This distribution elucidates the observed discrepancy where the White subgroup surpasses in accuracy, while the Black or African American subgroup excels in precision, recall, and F1 score metrics.

Table 1: Performance comparison between the subgroups of White and Black or African American.

Metric	**ResNet 3D**			**ResNet 3*2D**		
	White	B/AA	p	White	B/AA	p
Accuracy	0.844	0.776	<0.05	0.851	0.825	<0.05
Precision	0.678	0.812	<0.05	0.633	0.845	<0.05
Recall	0.494	0.580	<0.05	0.578	0.597	<0.05
F1 Score	0.572	0.677	<0.05	0.604	0.699	<0.05

Note: B/AA stands for Black or African American. p is the p value of difference tests.

Table 2: Confusion matrices for the ResNet 3*2D.

		Predicted	
		Positive	Negative
Actual	Positive	1,422	825
	Negative	1,038	8,985

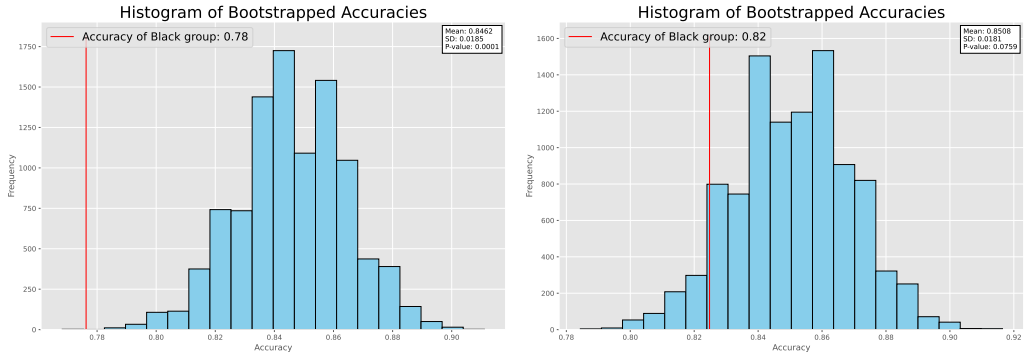
(a) White

		Predicted	
		Positive	Negative
Actual	Positive	71	13
	Negative	48	239

(b) Black or African American.

3.2. Single Evaluation Metrics

In this section, the accuracy is chosen as an example to show the impact of data uncertainty and the effectiveness of the proposed method. Although the disparity in performance metrics suggests that both models may exhibit fairness issues, the uncertainty of testing results needs to be further considered. To address this issue on uncertainty and rigorously assess the reliability of test results, bootstrapping techniques were applied to the data from the White group, using a sample size of 371 (matching the test sample size of the Black or African American group) and conducting 10,000 iterations. Given the extensive number of bootstrap samples, the resulting distribution of performance metrics is presumed to follow a normal distribution, allowing for the application of a z-test to evaluate the statistical significance of the observed performance disparities between the groups. The accuracy of each model was evaluated across the bootstrapped samples, with the distribution of these accuracy illustrated in histograms (referenced as Figure 2). The accuracy achieved on the Black or African American group is denoted by a



(a) ResNet 3D (b) ResNet 3*2D

Figure 2: Histograms of the bootstrapped samples from White group.The red line stands for the results from minority group. The further it is from the distribution of majority group, the larger probability the model has fairness problems..

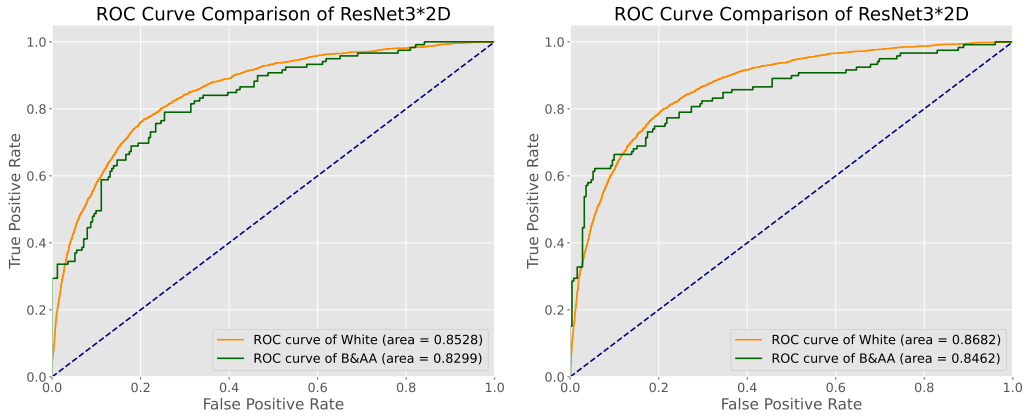
Table 3: The results of bootstrapped sample groups of majority groups and z-test between minority group and those sample groups.

Model	Mean Accuracy	Accuracy on Minority	p
ResNet 3D	0.846 ± 0.018	0.776	1e-4
ResNet 3*2D	0.851 ± 0.018	0.825	0.076

red line within these histograms. Utilizing the mean and standard deviation derived from these distributions, the p-value for the performance of the Black or African American group was calculated, with the outcomes detailed in Table 3. For a result to be considered statistically significant and indicative of a deviation from the expected distribution, the p-value must be lower than the threshold of 0.05. Based on this criterion, the ResNet 3D model is identified as having fairness issues due to its statistically significant deviation. In contrast, the results for the ResNet 3x2D model do not definitively confirm fairness concerns, as the p-value does not conclusively fall below the established threshold.

3.3. ROC Curve as Metrics

In this section, we demonstrate the application of our method using the ROC curve as the evaluation metric. Figure 3 displays the ROC curves for



(a) ResNet 3D

(b) ResNet 3*2D

Figure 3: ROC curves of both models on majority and minority groups.

both the White and the Black or African American subgroups. Notably, the ROC curves for the White group display a smoother trajectory compared to those of the Black or African American group, underscoring the increased uncertainty in the results for the Black or African American group due to its smaller sample size. For each bootstrapped sample, an ROC curve was generated and subsequently compared to the ROC curve of the Black or African American group using the ROC z-test. This comparison was structured around two null hypotheses: the first, denoted as "greater", posits that the model performs better on the bootstrapped White group samples than on the Black or African American group; the second, denoted as "less", assumes the inverse. The outcomes of these comparisons, including the count of significant p-values under each hypothesis and the p-value from a Binomial test (assuming a Null Hypothesis rate of 0.2), are summarized in Table 4. The results reveal that, for the "greater" hypothesis, the Binomial test p-values for both models are notably low, suggesting that a considerable proportion of the bootstrapped samples from the White group led to model performance that surpassed that of the Black or African American group. This finding underscores fairness issues within both models. Interestingly, while the accuracy metric alone did not conclusively reveal fairness concerns for the ResNet 3x2D model, the ROC curve analysis did. This discrepancy underscores the importance of using multiple evaluation metrics to provide a comprehensive assessment of a models fairness.

Table 4: Accuracy of different subgroups. "greater" stands for the number of significant p-values under hypothesis "greater", and "less" stands for the number of significant p-values under hypothesis "less".

Model	"greater"	"less"	p (Binomial test)
ResNet 3D	2,414	22	2.8e-24
ResNet 3*2D	2,360	22	4.7e-16

3.4. Ablation study

To validate the efficacy of the proposed evaluation methods, an ablation study was conducted focusing on the within-group variability for the White group, aimed at quantifying uncertainty. Initially, a subset of the White group, equivalent in size to the Black or African American group, was selected to serve as the "anchor" group. Subsequently, bootstrapping experiments were conducted comparing this "anchor" group against the remaining samples from the White group. Unlike the prior setup, the null hypothesis for these comparisons was modified to assess whether "the two ROC curves are significantly different". The frequency of significant p-values was tallied for each comparison. This procedure was iterated 100 times, with a new "anchor" group selected for each iteration, thereby generating a distribution of significant p-value counts. The distribution of these counts is depicted in Figure 4, illustrating the variability in significant p-values across the different "anchor" groups. The findings indicate that, for a majority of the "anchor" groups, the count of significant p-values was relatively low, suggesting minimal within-group variability. However, there were instances of "anchor" groups exhibiting an anomalously high number of significant p-values, potentially highlighting the presence of "hard cases" or data points that deviate substantially from the group's distribution [21, 22]. This analysis not only underscores the robustness and applicability of the proposed evaluation methods but also illustrates their potential utility in enhancing data curation processes, particularly in the identification and examination of challenging cases or outliers.

3.5. Effect on post processing mitigation methods

In this section, we discuss the impact of our proposed evaluation methods on existing bias mitigation techniques. Mitigation methods are commonly categorized into three groups: pre-processing, intra-processing, and

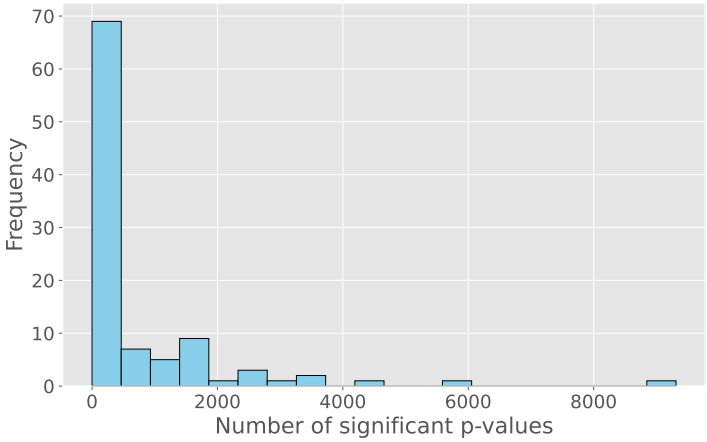


Figure 4: Histograms of the number of significant p-values for each "anchor" group.

post-processing [23]. Since our approach focuses on model evaluation without affecting the training process, it is naturally compatible with both preprocessing and intra-processing methods. However, post-processing methods, which directly influence model evaluation, may not integrate seamlessly with our techniques. To illustrate this, we conducted an experiment where we applied threshold adjustment [6], a widely implemented post-processing strategy, to balance accuracy and then assessed the performance gaps across various metrics. The basic idea is to modify the decision boundary or threshold used to classify instances. Usually, a model uses a probability threshold of 0.5 to decide between two classes. This threshold might be adjusted differently for different demographic groups to achieve fairness. However, when doing this, only one evaluation metric can be considered. As a result, while the gap of target metric might decrease with the new thresholds, the gap under other metrics might increase. In the experiment, we set accuracy as the metric to apply threshold adjustment, and calculated the performance gap under various evaluation metrics. The findings, presented in Table 5, show that while the accuracy gap between demographic groups was narrowed, the gaps under some other evaluation metrics might be enlarged. These results suggest that in scenarios with significant data imbalances across groups, preprocessing or intra-processing methods might be more effective than post-processing approaches.

Table 5: Performance comparison between the subgroups after threshold adjustment

Metric	**ResNet 3D**			**ResNet 3*2D**		
	White	B/AA	gap	White	B/AA	Gap
Accuracy	0.831	0.786	0.045 ↓	0.846	0.821	0.023 ↓
Precision	0.585	0.736	0.152 ↑	0.705	0.818	0.112 ↓
Recall	0.560	0.518	0.041 ↑	0.397	0.571	0.173 ↑
F1 Score	0.572	0.607	0.036 ↓	0.508	0.671	0.163 ↑

Note: Gap represents the absolute value of the difference between groups. ↓ means the gap is smaller after threshold adjustment, and ↑ means the gap is larger.

4. Discussion

The experiments conducted provide several critical insights. First, the choice of evaluation metrics significantly influences the determination of model fairness, highlighting the need for a comprehensive evaluation approach [24, 25]. In Sections 3.1 and 3.2, we illustrated how a model deemed fair under one metric may be identified as unfair under another. The ROC curve emerges as a robust option when a single evaluation metric is necessary, providing a balanced view of model fairness. Secondly, the importance of accounting for data uncertainty in the assessment of model fairness is highlighted. Specifically, the restricted sample size of testing data for minority groups can introduce performance discrepancies. Third, the ablation study suggests that minority groups could be considered as presenting challenging cases within the broader context of the majority group, a notion that finds relevance in clinical settings. The minor anatomical variations across different racial groups contribute to this scenario [26], emphasizing the critical need for the inclusion of high-quality medical data from diverse populations, particularly those representing minority groups. Further, given that our findings are based on a single dataset, we advocate for further validation of our methodology across a variety of datasets to confirm and possibly recalibrate the sample size considerations we have suggested. For now, we suggest that our evaluation method is particularly useful in scenarios where there is a noticeable disparity in sample sizes between groups.

5. Conclusions

While the majority of research in the domain of fairness has concentrated on addressing and mitigating bias within models, this paper posits that the evaluation of model fairness is equally crucial and presents its own set of challenges. The evaluation methodologies introduced herein effectively navigate through model biases and data uncertainties to provide a nuanced assessment of fairness across models. Despite the efficacy of these evaluation techniques, it is important to note that this study does not propose direct interventions for rectifying fairness issues. The task of completely neutralizing disparities in performance, especially given the significant variations within the dataset, remains a formidable challenge. Recently, the rapid advancement of generative models[27], such as diffusion models [28], offers promising avenues for addressing the limitations posed by data scarcity. Another potential approach would be to take multimodal information into account when develop AI models [29][30]. Nonetheless, achieving comprehensive fairness in machine learning models is an ongoing endeavor that requires continued effort and innovation.

Acknowledgement

This research was supported by National Science Foundation (2046708) (P.Y., principal investigator) and National Institutes of Health (R01EB032716) (G.W., contact principal investigator).

References

[1] H. Oh, C. Kim, Fairness-aware recommendation with meta learning, Scientific Reports 14 (1) (2024) 10125.

[2] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, F. Mahmood, Algorithmic fairness in artificial intelligence for medicine and healthcare, Nat Biomed Eng 7 (6) (2023) 719–742.

[3] J. Zhang, H. Chao, A. Dhurandhar, P.-Y. Chen, A. Tajer, Y. Xu, P. Yan, Spectral adversarial mixup for few-shot unsupervised domain adaptation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 728–738.

[4] X. Fang, P. Yan, Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, IEEE Transactions on Medical Imaging 39 (11) (2020) 3619–3629.

[5] G. Haskins, U. Kruger, P. Yan, Deep learning in medical image registration: a survey, Machine Vision and Applications 31 (1) (2020) 8.

[6] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 33233331.

[7] Q. Feng, M. Du, N. Zou, X. Hu, Fair machine learning in healthcare: A review, arXiv preprint arXiv:2206.14397 (2022).

[8] J. Zhang, H. Chao, A. Dhurandhar, P.-Y. Chen, A. Tajer, Y. Xu, P. Yan, When neural networks fail to generalize? a model sensitivity perspective, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 11219–11227.

[9] K. P. Seastedt, P. Schwab, Z. OBrien, E. Wakida, K. Herrera, P. G. F. Marcelo, L. Agha-Mir-Salim, X. B. Frigola, E. B. Ndulue, A. Marcelo, L. A. Celi, Global healthcare fairness: We should be sharing more, not less, data, PLOS Digital Health 1 (10) (2022) e0000102.

[10] A. Rajkomar, M. Hardt, M. D. Howell, G. S. Corrado, M. H. Chin, Ensuring fairness in machine learning to advance health equity, Annals of Internal Medicine 169 (2018) 866–872.

[11] L. Gao, L. Zhang, C. Liu, S. Wu, Handling imbalanced medical image data: A deep-learning-based one-class classification approach, Artificial Intelligence in Medicine 108 (2020) 101935.

[12] P. A. Lachenbruch, M. R. Mickey, Estimation of error rates in discriminant analysis, Technometrics 10 (1) (1968) 1–11.

[13] B. Efron, Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics 7 (1) (1979) 1 – 26.

[14] A. Isaksson, M. Wallman, H. Gransson, M. Gustafsson, Cross-validation and bootstrapping are unreliable in small sample classification, Pattern Recognition Letters 29 (14) (2008) 1960–1965.

[15] E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, Biometrics 44 (3) (1988) 837–845.

[16] U. Kaempf, The binomial test: a simple tool to identify process problems, IEEE Transactions on Semiconductor Manufacturing 8 (2) (1995) 160–166.

[17] J. Zhang, H. Chao, G. Dasegowda, G. Wang, M. K. Kalra, P. Yan, Revisiting the trustworthiness of saliency methods in radiology ai, Radiology: Artificial Intelligence 6 (1) (2023) e220221.

[18] H. Chao, H. Shan, F. Homayounieh, R. Singh, R. D. Khera, H. Guo, T. Su, G. Wang, M. K. Kalra, P. Yan, Deep learning predicts cardiovascular disease risks from lung cancer screening low dose computed tomography, Nature Communications 12 (1) (2021).

[19] K. Hara, H. Kataoka, Y. Satoh, Learning spatio-temporal features with 3D residual networks for action recognition, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 3154–3160.

[20] J. Zhang, H. Chao, X. Xu, C. Niu, G. Wang, P. Yan, Task-oriented low-dose ct image denoising, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24, Springer, 2021, pp. 441–450.

[21] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: International Conference on Learning Representations, 2017, p. arXiv:1610.02136.

[22] J. Zhang, H. Chao, P. Yan, Toward adversarial robustness in unlabeled target domains, IEEE Transactions on Image Processing 32 (2023) 1272–1284.

[23] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (jul 2021).

[24] J. Zhou, A. H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (5) (2021).

[25] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, et al., Metrics reloaded: recommendations for image analysis validation, Nature methods (2024) 1–18.

[26] D. L. Hongwei Hsiao, K. Snyder, Anthropometric differences among occupational groups, Ergonomics 45 (2) (2002) 136–152.

[27] I. Ktena, O. Wiles, I. Albuquerque, S.-A. Rebuffi, R. Tanno, A. G. Roy, S. Azizi, D. Belgrave, P. Kohli, T. Cemgil, A. Karthikesalingam, S. Gowal, Generative models improve fairness of medical classifiers under distribution shifts, Nature Medicine 30 (4) (2024) 1166–1173.

[28] F.-A. Croitoru, V. Hondru, R. T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (9) (2023) 10850–10869.

[29] P. Yan, G. Wang, H. Chao, M. K. Kalra, Multimodal radiology ai, Meta-Radiology 1 (2) (2023) 100019.

[30] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of chatgpt-related research and perspective towards the future of large language models, Meta-Radiology 1 (2) (2023) 100017.