Bias Mitigation for Medical Images

Wenhao Zhu ZWHSHY@YORKU.CA
EECS
York University
Toronto, Ontario, Canada

Chung-Yu Wang CYWANG14@YORKU.CA
EECS
York University
Toronto, Ontario, Canada

Alireza Daghighfarsoodeh ALIREDAQ@YORKU.CA
EECS
York University
Toronto, Ontario, Canada

Abstract

The advancement of Artificial Intelligence (AI) provides a convenient life for people but it also prompts concerns about the equality of human rights. Doubts linger as to whether individuals, regardless of background, will be afforded equal opportunities to access fundamental rights such as education and medical resources. Evidence reveals disparities in false positive rate (FPR) across various subgroups, casting shadows on the fairness of AI outcomes. In this project, we aim to diminish bias within the AI model through undersampling and upsampling techniques. We discovered a model that could mitigate the bias among various subgroups. Our code and results are available at https://github.com/wangjohn5507/Fairness_project.git.

1. Introduction

Recently, the development of Artificial Intelligence (AI) has greatly enhanced various aspects of the quality of human's life, such as autonomous vehiclesMuthalagu et al. (2021), image classificationLu and Weng (2007), spam detectionCrawford et al. (2015) etc. However, there is also a growing concern among the public regarding trustworthiness of AI systems. For example, Flores et al. (2016) demonstrated that there existed unfairness between white and black people for providing different failure rates in a risk assessment system (known as COMPAS). Moreover, Seyyed-Kalantari et al. (2021) illustrate different false positive rate (FPR) gap among different subgroups by using AI to predict whether people have specific diseases based on their chest X-Ray. All those cases display the potential unfairness in AI systems. Consequently, this draws more attention to mitigation of the unfairness in AI among different protected groups, especially in the medical field where differential treatment among subgroups could have significant consequences.

Although there are a few approaches that researchers have already applied to mitigate the unfairness in the field of medical image, Chouldechova (2017) proposed the impossibility

theorem which stated that we cannot achieve both equal false positive rate (FPR) and equal false negative rate (FNR) across different subgroups when they have different prevalence. Hence, we should only focus on only one of these two fairness notation.

Also, in the Machine Learning (ML) pipeline, several aspects can contribute to the unfairness of machine learning systems, such as the dataset, preprocessing, in-model training, etc. To address unfairness at different stages of the Machine Learning pipeline, various approaches should be applied to mitigate these issues.

As illustrated previously, existing Machine Learning systems exhibit varying false positive rate gaps among different subgroups. The unfair treatment of different subgroup can exacerbate social disparity problems, particularly within the medical field. Therefore, mitigating unfairness in this domain is of great importance. This paper specifically focuses on addressing mitigation of fairness in the AI models which are used to predict the diseases of the patients by their chest radiographs.

To sum up, our main contributions are the following:

- Proposed data preprocessing techniques to mitigate the bias in the medical dataset
- Demonstrated the ablation study of mitigating bias based on different Loss functions

2. Related Work

Seyyed-Kalantari et al. (2020). examine state-of-the-art deep neural classifiers on large public medical imaging datasets, focusing on fairness across different subgroups of protected attributes. They analyze true positive rate (TPR) disparities among patient sex, age, race, and insurance type, revealing extensive patterns of bias in classifiers across all datasets. Their findings highlight the critical importance of auditing for algorithmic disparities to ensure equitable clinical decision-making as models transition from research to practical applications.

The study conducted by Seyyed-Kalantari et al. (2021). systematically examines the issue of underdiagnosis bias in AI-based chest X-ray (CXR) prediction models across multiple large public radiology datasets. It highlights how such models, despite their advanced capabilities, consistently underdiagnose historically underserved patient populations, including those differentiated by race, socioeconomic status, sex, and age. The investigation reveals a higher rate of underdiagnosis for intersectional subpopulations, such as Hispanic female patients, emphasizing the potential for AI systems to exacerbate existing healthcare disparities if deployed without addressing these biases. This research underscores the ethical and clinical implications of implementing AI diagnostic tools without rigorous bias mitigation strategies.

Zhang et al. (2023) propose a novel framework aimed at mitigating biases in machine learning models through a debiasing training strategy. The core of their methodology revolves around the identification and subsequent neutralization of bias in data prior to the training phase, thereby ensuring that the resultant models exhibit fairer outcomes across diverse demographic groups. By integrating bias identification directly into the training process, their approach seeks to address the root causes of bias in algorithmic decision-making, distinguishing it from post-hoc adjustment strategies that attempt to correct bias after a model has been trained. The authors demonstrate the effectiveness of their technique through a series of experiments, showcasing significant improvements in fairness metrics

without compromising the overall accuracy of the models. This work contributes to the growing body of research on ethical AI, offering a practical solution for developers and practitioners aiming to build more equitable machine learning systems.

In their investigation on improving fairness in chest X-ray classifiers, Zhang et al. (2022) explore various methods, notably including Empirical Risk Minimization (ERM) and its variants—Balanced ERM and Stratified ERM—for baseline comparisons. They also evaluate methods aimed at achieving group fairness, such as adversarial approaches, MMDMatch, MeanMatch, and FairALM, which attempt to enforce fairness constraints during model training. Furthermore, they assess methods focused on improving the performance of the worst-case group, like GroupDRO and ARL, which adjust training emphasis to enhance outcomes for groups traditionally disadvantaged by model biases. Their findings suggest that simple data balancing techniques often perform comparably to more complex debiasing methods, highlighting the challenges and nuances associated with integrating fairness into clinical diagnostic tools.

Zhang et al. (2022) investigates the application of algorithmic fairness in chest X-ray diagnosis through deep learning models. Their method involves evaluating a state-of-the-art chest X-ray classifier for fairness, using algorithmic interventions to rectify identified biases. They critically examine the reasons behind the model's unfairness and the serious side effects of algorithmic approaches to achieve equal predictive performance across demographic groups. Their findings advocate for a nuanced understanding of fairness, emphasizing the importance of addressing biases in data over applying algorithmic solutions indiscriminately.

Glocker et al. (2023) explore how deep learning models applied to chest X-ray disease detection may inadvertently encode and utilize protected characteristics like race and sex, potentially due to biases in the training data. The study employs a multifaceted methodological approach, including test-set resampling, transfer learning, multitask learning, and model inspection, to evaluate and understand performance disparities across demographic subgroups. This comprehensive analysis aims to uncover the extent to which protected characteristics influence model performance, highlighting the critical need for fairness and equity in medical diagnostic algorithms.

3. Dataset

In this paper, we would like to use the MIMIC-CXR Dataset. Sellergren et al. introduced an image embedding for the dataset of MIMIC V2.0.0 Chest X-Ray database. This dataset is available on physionet and it required the completion of CITI certificate. The entire dataset contains around 370 thousand chest radio graphs. We will conduct our research based on four different sensitive attributes (Race, Gender, Age and Insurance).

3.1. Race

As shown in the figure 1, the dataset is predominantly White, with all other racial groups collectively making up less than 20% of the data. Specifically, the American Indian/Alaska Native category comprises just 384 entries, representing 0.3% of the total dataset.

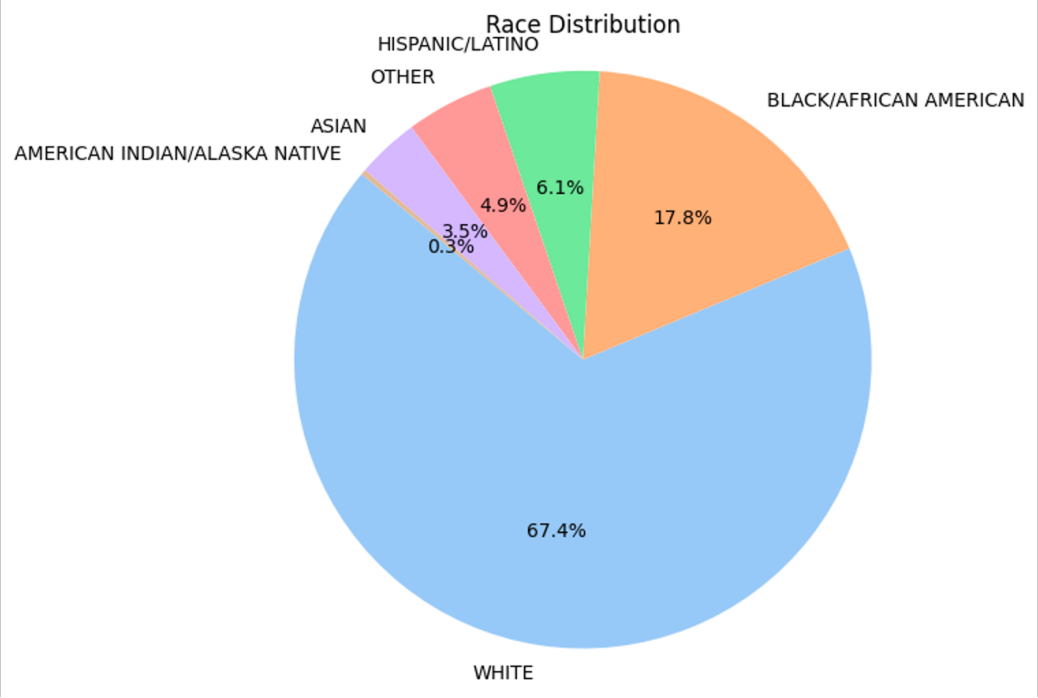

Figure 1: Race distribution of the dataset.

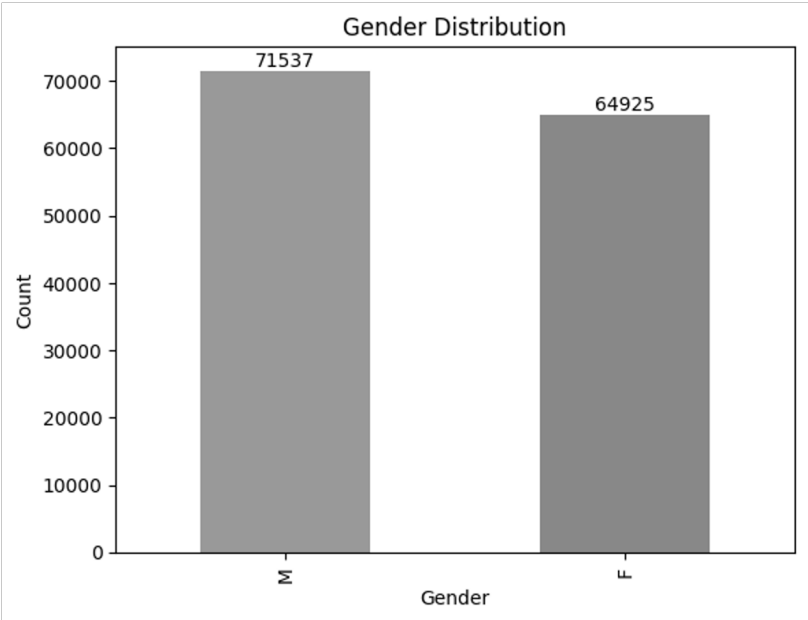

Figure 2: Gender distribution of the dataset.

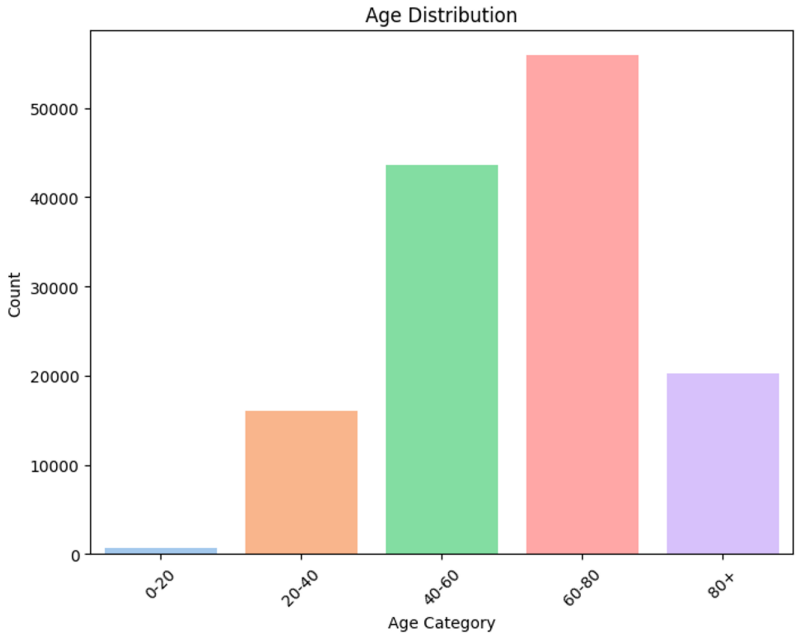

Figure 3: Age distribution of the dataset.

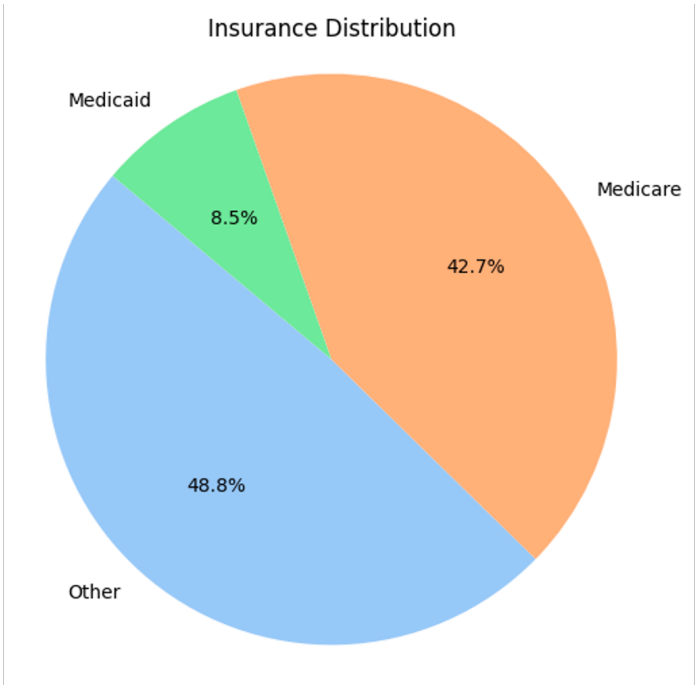

Figure 4: Insurance distribution of the dataset.

3.2. Gender

As shown in the figure 2, more than half of the data consist of men, and even this slight majority contributes to bias in the model.

3.3. Age

As shown in the figure 3, the majority of the dataset consists of individuals aged between 60 and 80 years. People aged 0-20 years are less likely to undergo chest X-rays, indicating a lower representation of this age group in the dataset.

3.4. Insurance

As shown in the figure 4, the majority of the dataset falls under the "Other" category for insurance at 48.8%. Medicaid accounts for only 8.5% of the dataset, indicating a significant bias in the distribution of insurance types represented.

4. Methodology

The methodology section of the paper describes initiating the process by applying Generalized Cross Entropy (GCE) Loss Nam et al. (2020) to train the dataset in conjunction with a classifier, with meticulous monitoring of the loss associated with each data point. This step forms the basis for the initial model. Utilizing the detailed loss data, the researchers identify and isolate biased data points within the dataset. These data points are subsequently targeted for upsampling and downsampling, processes designed to debias the models. This procedure includes retraining the previously developed model on the adjusted dataset to recover accuracy while striving to eliminate bias. The researchers' method incorporates both the elimination and adjustment of biased data with strategic retraining techniques such as upsampling and downsampling, establishing a comprehensive approach to mitigating bias in machine learning models.

4.1. Train Classifier with GCE Loss

In the study, the objective is to minimize the Generalized Cross Entropy (GCE) Loss to identify images that most closely align with biases present within the dataset. By training the original model using GCE Loss as a classifier, it is possible to extract biased images in an unsupervised manner An et al. (2022). Consequently, the classifier was trained utilizing GCE Loss. The GCE Loss can be defined as follows for a single prediction:

$$L_{GCE}(p(x;\theta)) = -\frac{1 - p_y(x;\theta)^q}{q} \tag{1}$$

where $p(x;\theta)$ denotes the softmax results from a classifier θ, $p_y(x;\theta)$ indicates the probability assigned to the target variable y with $q \in (0, 1]$ which is the degree of controlling amplification.

The training classifier details involve initially inputting the dataset with the image embedding vector D, which contains images X that potentially biases. A classifier, denoted as θ, is then trained on D using Generalized Cross Entropy (GCE) Loss. This classifier is designed to predict the presence of features in images that align with biases. After the model has been trained using GCE Loss, it is utilized to produce a loss value, $L_{GCE}(p(x;\theta))$, for each image, indicating the degree of alignment of each image x_i with the identified biases. Once loss values are assigned to each image, the dataset is sorted in ascending order based on these loss values. Following this, upsampling and downsampling strategies are

employed, utilizing the sorted dataset to effectively manage the distribution of images in further analyses.

4.2. Upsampling Strategy

The primary objective of the upsampling strategy is to enhance the model's capability to learn more features from the most biased data Fattal (2007). This strategy focuses on manipulating the dataset by duplicating the top $K\%$ of data Dai et al. (2021) that exhibits the lowest Generalized Cross Entropy (GCE) Loss values, which are indicative of high bias alignment. By selecting and amplifying this segment of the dataset, the upsampling approach increases the prevalence of the most biased features within the training set. Consequently, the model is repeatedly exposed to these biased features, which aims to improve its proficiency in recognizing similar patterns in new data. This method not only enhances the model's exposure to critical biased data but also seeks to optimize its performance in environments where these biased features are prevalent.

Specifically, the proportion of each subgroup was evaluated to determine if its False Positive Rate (FPR) Chaipanha and Kaewwichian (2022) exceeded the overall FPR. If this condition was met, upsampling was applied to the subgroup by duplicating the top $K\%$ of data with low Generalized Cross Entropy (GCE) Loss. Here, 'K' represents the hyperparameter of the approach.

4.3. Downsampling Strategy

The primary objective of the downsampling strategy is to reduce the model's exposure to highly biased data, thereby encouraging it to learn fewer features from these data points. To implement this strategy, the approach involved removing the top $K\%$ of data characterized by low Generalized Cross Entropy (GCE) Loss, which signifies a high degree of bias alignment. This selective removal of the most biased data aims to diminish the impact of these features on the model's training process. By reducing the representation of biased data within the training set, the model is less likely to overfit these characteristics and can develop a more generalized understanding of the dataset. This downsampling technique plays a crucial role in attempting to balance the dataset and improve the model's overall performance by focusing on less biased learning.

In implementing the downsampling strategy, an initial assessment of each subgroup was conducted to determine its proportion relative to the entire dataset. If a subgroup constituted more than $N\%$ of the total dataset and its False Positive Rate (FPR) was lower than the overall FPR, it was identified as a candidate for downsampling. The base weight for this process was set at $K\%$, which served as a starting point for further adjustments. 'N' and 'K' are hyperparameters of the approach. To refine the amount of data to be removed, two sigmoid-like functions were employed. These functions, in conjunction with the maximal distance between FPR of different subgroups and their respective proportions within the dataset, served as critical inputs in determining the precise size of the dataset to be downscaled. This approach allowed for tailored adjustments, ensuring that downsampling was both effective in reducing bias and sensitive to the dynamics within different subgroups, thereby optimizing the overall training process of the model.

5. Experiment Design

Model Description

We employed a single-layer deep neural network specifically optimized for multimodal prediction tasks within the domain of disease forecasting. This architectural design facilitates the integration of various data modalities, aiming to significantly improve the model's predictive accuracy and reliability. The training process of the model extended over 50 epochs with a batch size of 32, which helps in stabilizing the learning curve. Furthermore, the neural network features a hidden layer comprising 196 units, which is configured to process and synthesize the information efficiently from the diverse data sources involved. This structure is crucial for capturing complex patterns that are essential for accurate disease prediction.

Loss Calculation and Bias Mitigation

Our methodology focuses primarily on the application of Generalized Cross Entropy (GCE) Loss to recalibrate and optimize our neural network's predictive capabilities.

GCE Loss

GCE Loss is employed as the principal loss function for training the model. It calculates loss more flexibly by lowering the penalty on well-classified examples, which helps in handling noisy labels and reducing model bias. Each datum's loss is recalculated, and these values are stored for subsequent analysis aimed at bias mitigation.

Ablation Study: BCE Loss

In an ablation study, we utilize Binary Cross-Entropy (BCE) Loss to contrast the effects and efficiencies relative to GCE Loss. This comparative analysis serves to underline the enhancements in prediction and bias reduction afforded by the GCE Loss approach.

Sampling Techniques for Bias Correction

To further address potential biases in prediction, we implement both down-sampling and up-sampling strategies based on computed loss values:

Down Sampling: We set 'K' as 5, meaning that 5% of top Instances demonstrating lower loss measurements and a False Positive Rate (FPR) below the overall average are selectively excluded. Also, we set 'N' as 15, meaning that the proportion of the subgroup should be more than 15% in total to remove. This strategy aims to diminish the influence of outliers or underperforming data points on the learning algorithm.

Up Sampling: Conversely, data points characterized by higher losses and a higher FPR than average are augmented. We experimented with increasing the dataset by 20%, 25%, and 50%, to assess the impacts on the model's performance and bias.

Integrated Sampling Strategy

Combining both sampling approaches, we investigate their synergistic effects on enhancing model precision and reducing biases. This integrated sampling framework is pivotal in our comprehensive evaluation of the model's performance across varied operational scenarios.

Once we have developed the debiased model, we will examine its classification outcomes using the False Positive Rate (FPR) metric and contrast these findings with those from the original model. An improved FPR across various subgroups signifies enhanced fairness of our model. Additionally, we will assess and compare the accuracy of both models to evaluate the extent to which the debiasing process impacts the overall performance of the model, thereby understanding the trade-off between debiasing and model accuracy.

Table 1: Multiple Methods On GCE Loss

No Finding	DS	US 20%	US 25%	US 50%	D& U 20%	D& U 25%	D& U 50%	Before
AUC	85 ↓1	85 ↓1	85 ↓1	85 ↓1	84 ↓2	84 ↓2	86 -	86
Gender FPR Gap	0.025 ↑3.8%	0.022 ↓8.1%	0.021 ↓11.1%	0.023 ↓3.7%	0.019 ↓19.4%	0.027 ↑11.8%	0.02 ↓17.5%	0.024
Race FPR Gap	0.206 ↓4.8%	0.22 ↑1.3%	0.19 ↓12.1%	0.188 ↓13.3%	0.166 ↓23.1%	0.182 ↓15.7%	0.182 ↓15.7%	0.217
Insurance FPR Gap	0.106 ↓15.1%	0.098 ↓21.5%	0.098 ↓21.5%	0.113 ↓9.9%	0.098 ↓40.4%	0.095 ↓23.9%	0.105 ↓16.1%	0.125
Age FPR Gap	0.147 ↓10.9%	0.119 ↓27.4%	0.114 ↓30.6%	0.13 ↓19.9%	0.099 ↓20.8%	0.104 ↓37%	0.121 ↓36.3%	0.165

Table 2: Upsampling for different groups with GCE Loss

No Finding	Race	Gender	Insurance	Age	Before
AUC	85 ↓1	85 ↓1	85 ↓1	85 ↓1	86
Gender FPR Gap	0.025 ↑3.6%	0.029 ↑20.1%	0.025 ↑3.4%	0.023 ↓2.9%	0.024
Race FPR Gap	0.209 ↓3.5%	0.196 ↓9.6%	0.209 ↓3.5%	0.182 ↓15.7%	0.217
Insurance FPR Gap	0.102 ↓18.2%	0.099 ↓20.3%	0.109 ↑12.5%	0.095 ↓23.6%	0.125
Age FPR Gap	0.113 ↓31.6%	0.118 ↓28.5%	0.123 ↓25.3%	0.123 ↓25.5%	0.165

Table 3: Multiple Methods On BCE Loss

No Finding	DS	US 20%	US 25%	US 50%	D& U 20%	D& U 25%	D& U 50%	Before
AUC	85 ↓1	86	84 ↓1	86	86	86	86	86
Gender FPR Gap	0.009 ↓60.3%	0.02 ↓15%	0.025 ↑4.4%	0.019 ↓18.5%	0.026 ↑6.2%	0.026 ↑6.2%	0.022 ↓6.7%	0.024
Race FPR Gap	0.216 ↓0.8%	0.209 ↓3.5%	0.175 ↓19%	0.214 ↓1.2%	0.267 ↑23.3%	0.225 ↑3.7%	0.256 ↑18.3%	0.217
Insurance FPR Gap	0.102 ↓18.2%	0.127 ↑1.8%	0.128 ↑2.3%	0.117 ↓6.4%	0.136 ↑8.5%	0.133 ↑6.3%	0.117 ↓6.2%	0.125
Age FPR Gap	0.153 ↓6.8%	0.205 ↑24%	0.201 ↑21.9%	0.178 ↑8.1%	0.183 ↑11.3%	0.137 ↓4.9%	0.175 ↑6.4%	0.165

6. Result and Discussion

Effectiveness of GCE Loss with Combined Sampling Strategies

As depicted in table 1, employing GCE Loss with a combination of 50% up-sampling and down-sampling consistently narrows the False Positive Rate (FPR) gap. Importantly, this is achieved without a significant drop in the Area Under the Curve (AUC). As shown in figure 6, we narrow the gap with GCE Loss. On average, this approach leads to a 21.4% reduction in the FPR gap, marking it as the most effective strategy for mitigating bias within our GCE Loss framework.

Table 4: Upsampling for different groups with BCE Loss

No Finding	Race	Gender	Insurance	Age	Before
AUC	86	86	86	86	86
Gender FPR Gap	0.022 ↓6.3%	0.023 ↓3.1%	0.023 ↓3.2%	0.018 ↓25.5%	0.024
Race FPR Gap	0.238 ↑9.8%	0.196 ↓9.6%	0.235 ↑8%	0.211 ↓2.4%	0.217
Insurance FPR Gap	0.135 ↑7.5%	0.122 ↓2.1%	0.129 ↑2.9%	0.112 ↓10.4%	0.125
Age FPR Gap	0.18 ↑9.3%	0.174 ↑5.4%	0.178 ↑8.5%	0.17 ↑3.1%	0.165

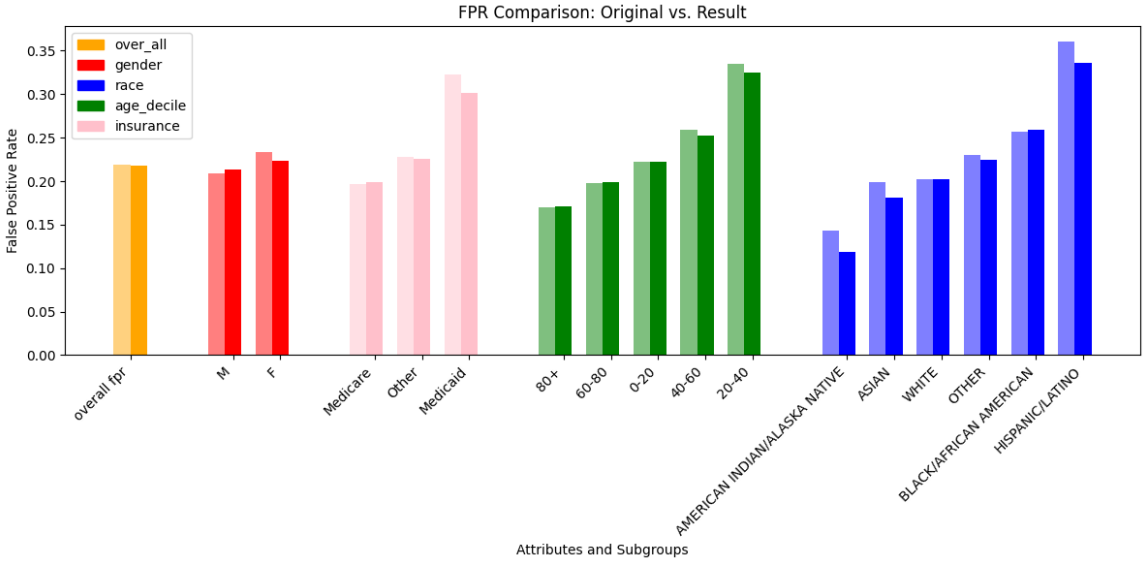

Figure 5: Optimal outcomes were achieved using Binary Cross-Entropy Loss. This method consistently facilitated a reduction in the False Positive Rate and narrowed the performance gap.

Impact of GCE Loss with Singular Sensitive Attribute Focus

Analysis presented table 2 shows that targeting a single sensitive attribute with GCE Loss also results in a consistent reduction in the FPR gap. Focusing solely on age proves most advantageous, reducing the gap by an average of 17%. This targeted approach does not significantly compromise the AUC, illustrating its efficacy.

Performance of BCE Loss with Down-Sampling

As shown in table 3 and further illustrated in figure 5, implementing BCE Loss with down-sampling significantly improves the FPR across all sensitive groups. Since some of the results of table 3 show that some gaps are increasing, it is not promising to mitigate the FPR gap across all sensitive groups. However, this strategy reaches the best result of mitigating the bias across all the experiments, resulting in a substantial average reduction of 21.5% in the FPR gap, highlighting its effectiveness in promoting fairness in predictions.

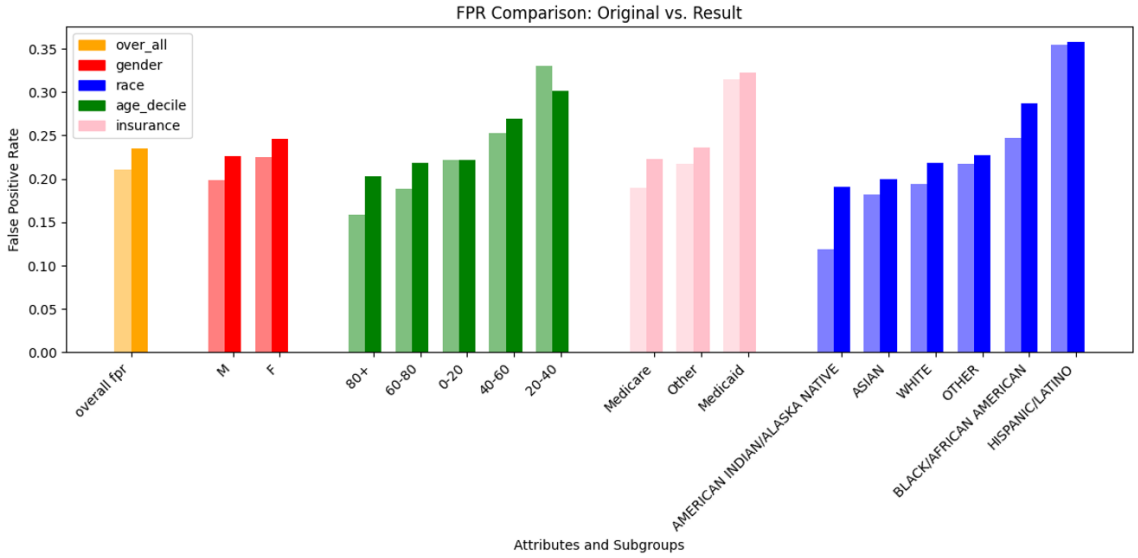

Figure 6: Optimal outcomes were observed with Generalized Cross-Entropy Loss. This approach routinely contributed to a narrowing of the discrepancy.

Focusing on a Single Sensitive Attribute with BCE Loss

Results from table 4 indicate that using BCE Loss to focus on a single sensitive attribute, such as age, effectively reduces the FPR gap. This method results in a moderate reduction, averaging a 4.5% decrease in the FPR gap, while maintaining stable AUC metrics. This suggests a balanced approach to reducing bias without sacrificing predictive performance.

7. Limitation and Future Work

7.1. Limitation

Our research has identified two methods that outperform others within their respective loss functions. While these findings underscore the effectiveness of specialized approaches, they also prompt us to scrutinize the universality of our own method. In light of this, we have identified several limitations inherent in our approach.

Schroter et al. (2004) showed **the quality of data set** emerges as a significant limitation influencing the performance of the project. In our methodology, our aim is to identify instances from the training data with the top k lowest Generalized Cross Entropy (GCE) Loss, which reveals bias inherent in the data set. By eliminating these samples, we strive to alleviate bias within the training data set, thus fostering a fairer model. However, this endeavor is inevitably constrained by the quality of the data set itself. Consider a scenario where high-quality samples are scarce. Removing too many useful training samples could result in a decline in the overall performance of the model. Therefore, the quality of the data set has a huge influence over the efficiency of our approach.

Secondly, another potential limitation to our approach lies in **data shits and domain drifts**. Our project primarily concentrates on the embedding data set of chest radiography from the United States. It remains uncertain whether the effectiveness of our approach can be replicated in other regions, such as Asia or Africa. Furthermore, our focus solely on the chest radiography embedding data set raises questions about the applicability of our approach to images from other fields, such as Magnetic resonance imaging (MRI) scans , computed tomography (CT) Scans or pathology images. Therefore, it is imperative to conduct further studies to validate the generalizability of our method beyond the specific dataset and geographical context we have examined.

Lastly, in our approach, we did not explore **alternative debiasing methodologies**. Our primary focus revolves around mitigating biases within the data set itself, which can be considered as pre-processing in machine learning. However, within the machines learning pipeline, there exist several ways to explore, including aspects like in-model training or post processing techniques. Thus, relying solely on pre-processing the dataset may not suffice to yield promising results.

7.2. Future Work

Due to the limitations of time and resources, there are several avenues for future improvements to enhance the robustness of our results.

Firstly, our experiment only encompasses two loss functions. However, there are a few loss function available for further exploration. Future endeavors could involve investigating **additional loss functions** and even combinations thereof to potentially yield superior results.

Secondly, our experiment is confined to a single data set, which is MIMIC-CXR. Broadening the scope to encompass **multiple data sets** could provide valuable insights into the generalizability and applicability of our approach across diverse data sources.

Lastly, our results are derived from a single execution of our method. Employing a **5-fold validation** approach on the data set in future iterations would bolster the reliability and robustness of our findings. this cross-validation technique would provide a more comprehensive assessment of the consistency and stability of our method's performance across varying data sub sets.

8. Conclusion

In this project, we provide a promising approach to upsampling and downsampling the training data by GCE Loss function and BCE loss function , which provide two outstanding methods, which mitigate the biases among different subgroups. However, it does worth future study to employ other loss functions with our current method.

References

Jaeju An, Youngsang Kwak, and Jaekwang Kim. Mitigating dataset bias via image translation. In *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS)*, pages 1–7. IEEE, 2022.

Wuttikrai Chaipanha and Patiphan Kaewwichian. Smote vs. random undersampling for imbalanced data-car ownership demand model. *Komunikácie*, 24(3), 2022.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Michael Crawford, Taghi M Khoshgoftaar, Joseph D Prusa, Aaron N Richter, and Hamzah Al Najada. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2:1–24, 2015.

Yutong Dai, Hao Lu, and Chunhua Shen. Learning affinity-aware upsampling for deep image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6841–6850, 2021.

Raanan Fattal. Image upsampling via imposed edge statistics. In *ACM SIGGRAPH 2007 papers*, pages 95–es. 2007.

Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *EBioMedicine*, 89: 104467, March 2023.

Dengsheng Lu and Qihao Weng. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5): 823–870, 2007.

Raja Muthalagu, Anudeep Sekhar Bolimera, Dhruv Duseja, and Shaun Fernandes. Object and lane detection technique for autonomous car using machine learning approach. *Transport and Telecommunication Journal*, 22(4):383–391, 2021.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

Sara Schroter, Nick Black, Stephen Evans, James Carpenter, Fiona Godlee, and Richard Smith. Effects of training on quality of peer review: randomised controlled trial. *Bmj*, 328(7441):673, 2004.

Andrew Sellergren, Atilla Kiraly, Tom Pollard, Wei-Hung Weng, Yun Liu, Akib Uddin, and Christina Chen. Generalized image embeddings for the mimic chest x-ray dataset.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y-Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. pages 232–243, 2020.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Robert Pfohl, and Ghassemi Ghassemi. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. March 2022.

Haoran Zhang, Thomas Hartvigsen, and Marzyeh Ghassemi. Algorithmic Fairness in Chest X-ray Diagnosis: A Case Study. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, (Winter 2023), feb 27 2023. https://mit-serc.pubpub.org/pub/algorithmic-chest.