arXiv:2412.00606v3 [cs.AI] 18 Sep 2026

Fairness at Every Intersection: Uncovering and Mitigating Intersectional Biases in Multimodal Clinical Predictions

Kishore Sampath[1],**Ayaazuddin Mohammad**[1], **Resmi Ramachandranpillai**[2],

[1]Northeastern University, USA
[2]Institute for Experiential AI, Northeastern University, USA
{sampath.ki, mohammad.ay, r.ramachandranpillai}@northeastern.edu

Abstract

Biases in automated clinical decision-making using Electronic Healthcare Records (EHR) impose significant disparities in patient care and treatment outcomes. Conventional approaches have primarily focused on bias mitigation strategies stemming from single attributes, overlooking intersectional subgroups - *groups formed across various demographic intersections (such as race, gender, ethnicity, etc.)*. Rendering single-attribute mitigation strategies to intersectional subgroups becomes statistically irrelevant due to the varying distribution and bias patterns across these subgroups. The multimodal nature of EHR - *data from various sources such as combinations of text, time series, tabular, events, and images* - adds another layer of complexity as the influence on minority groups may fluctuate across modalities. In this paper, we take the initial steps to uncover potential intersectional biases in predictions by sourcing extensive multimodal datasets, MIMIC-Eye[1] and MIMIC-IV ED, and propose mitigation at the intersectional subgroup level. We perform and benchmark downstream tasks and bias evaluation on the datasets by learning a unified text representation from multimodal sources, harnessing the enormous capabilities of the pre-trained clinical Language Models (LM), MedBERT, Clinical BERT, and Clinical BioBERT. Our findings indicate that the proposed sub-group-specific bias mitigation is robust across different datasets, subgroups, and embeddings, demonstrating effectiveness in addressing intersectional biases in multimodal settings.

1 Introduction

Electronic Healthcare Records (EHR) have revolutionized the optimization and delivery of healthcare, providing a digital database of patient information that enhances the efficiency and effectiveness of clinical decision-making (Seymour, Frantsvog, and Graeber 2012). Automating clinical decision-making using EHR is not without significant challenges, particularly with the perpetuation and amplification of biases. These biases can adversely impact the quality and equity of care provided to diverse patient populations and their intersections (Pivovarov et al. 2014; Boyd et al. 2023; Rouzrokh et al. 2022; Banerjee et al. 2023).

Existing measures for intersectional biases in healthcare (Ogungbe, Mitra, and Roberts 2019; Okoro, Hillman, and Cernasev 2022) overlooked the multimodal dimensionality (Turan et al. 2019; Amal et al. 2022; Kline et al. 2022; Pang et al. 2021) as learning from diverse sources affects data coherence, feature interoperability, and introduces varying bias patterns. Some efforts have been made to learn unified feature representation for disease diagnosis generation (Niu et al. 2024) and emergency department predictions (Lee et al. 2024). Figure 1 shows various challenges in dealing with intersectional biases in multimodal learning. The challenge of intersectionality becomes even more pronounced in multimodal settings as the fairness value fluctuates with different modality information. Moreover, single-attribute based fairness mitigation strategies become irrelevant as the subgroup densities vary significantly when considering demographic intersections, and balancing those sub-group populations in multimodal training (such as through counterfactual generation and data reweighting) is not practically feasible (Alabdulmohsin et al. 2024). In Figure 1, training is conducted with bias mitigation applied separately for race and gender. During model inference (Case 1), the model exhibits a significant bias toward the intersectional patient group formed by race and gender, utilizing information from all available

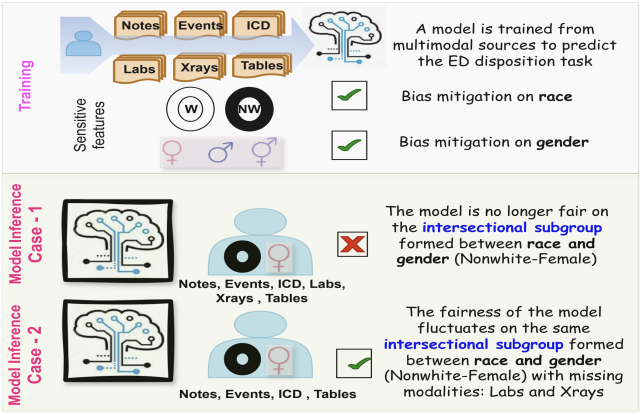


Figure 1: Intersectional biases in multimodal settings; Bias mitigation focused at single attribute is no longer valid for demographic intersections (case 1). The nature of multimodal data adds another layer of complexity in terms of fairness fluctuations (case 2).

[1]The only multimodal dataset available for research that encloses patients' information from all modalities (including x-rays).

data sources. However, in Case 2, the same model does not show bias toward the intersectional patient group when X-rays and lab results are excluded. It is thus important to note that modality-specific bias mitigation will fail in this scenario as different modalities exhibit different bias patterns and intensities (Yan, Huang, and Soleymani 2020)[2] and the biases in multimodal model may not correlate with those found in individual modalities.

Motivated by the aforementioned challenges, we propose **Subgroup-specific Discrimination Aware Ensembling (SDAE), an ensemble of multimodal classifier models tailored to address layered complexities of demographic intersectional biases**. **To capture vast knowledge of patient's health conditions from multimodal sources, we rely on the extensive power of pre-trained LMs: MedBERT (Rasmy et al. 2021), Clinical BERT, and BioClinical BERT (Alsentzer et al. 2019) for generating clinically relevant embeddings in the form of texts**. **Employing pre-trained clinical LMs to generate embeddings has an added advantage of facilitating learning with small-scale data, like MIMIC-Eye, harnessing the potential of rich and high dimensional clinical knowledge acquired through pre-training**. To accommodate these clinical LMs, we learn the **representations in unified textual format from all diverse multimodal sources**. We perform evaluations using MIMIC-Eye (Hsieh et al. 2023) and MIMIC-IV ED (Johnson et al. 2020) and our method is robust to any tasks/domains involving demographic intersections and multimodal learning.

Note that the MIMIC-IV ED does not exhibit any severe disparities in terms of the intersections. Therefore, to test the robustness of our method, we created a biased sample[3] from MIMIC-IV ED and used it for all the downstream tasks and bias analysis.

Contributions. Our major contributions address the complex challenges of intersectional biases in multimodal healthcare. Additionally, the MIMIC-Eye dataset has not yet been explored for downstream tasks and biases due to its inherent complex nature of multimodality which amplifies the overall impact of our work. The major contributions can be summarized as:

1. We propose bias mitigation through multimodal subgroup-specific ensemble classifiers. We design a post-process bias mitigation customized at the sensitive subgroup intersections (Section 3.3).
2. We benchmark downstream tasks on MIMIC-Eye analyzing modality-wise contributions to prediction accuracy and comparing these results to the overall prediction performance achieved through multimodal learning (Section 4.1).
3. In-depth intersectional bias analysis on the benchmarking tasks on MIMIC-Eye (Section 4.2).
4. Empirical analysis of single attribute post-process bias mitigation approaches and their effect on intersectionality. We show that existing bias mitigation approaches cannot be extended to settings combining intersectionality and multimodality (Section 4.2).

[2]The authors have proposed data balancing and adversarial strategies, but the multimodality dimension is limited.

[3]More details and original bias evaluation are given in the supplementary file, Section 5

2 Background

We consider binary and multitask (binary) classifications. The data, $D = \{X_n\}_{i=1}^{n}$ of each instance contains information from clinical notes (EHR_notes_n), Xrays/radiology images (EHR_Xray_n), clinical events (EHR_events_n), laboratory test results (EHR_lab_n), and from the tabular structure ($EHR_structured_n$). The sensitive features S are from the $EHR_structured_n$. The data instance X encloses information from all the modalities and can be written as:

$$X_i = \begin{pmatrix} EHR_notes_i, EHR_events_i, \\ EHR_lab_i, EHR_Xray_i, \\ EHR_structured_i \end{pmatrix} \quad (1)$$

The downstream prediction tasks on each X_i contain a ground truth label $Y \in \{0, 1\}$. Since we consider healthcare settings, we don't differentiate between favorable and unfavorable classes, as it depends on the underlying task at hand. We assume multiple sensitive settings, $s = \{s_1, s_2, ...s_m\}$, where m represents the total number of sensitive attributes in the dataset. Then we define subgroups, $\text{SG}_{gs_1 \times gs_2 \times ... gs_m} = gs_1 \cap gs_2 \cap ... \cap gs_m$ who belong to group gs_1 through gs_m with marginal sensitive attributes $s_1, s_2, ...s_m$. For example, if $s = \{gender, race\}$, $gender \in \{male, female\}$, and r$ace \in \{white, non - white\}$, then we will have 4 subgroups, SG, (i) $\{female, white\}$, (ii) $\{female, non{-}white\}$, (iii) $\{male, white\}$, and (iv) $\{male, non{-}white\}$, |SG| = 4.

2.1 Algorithmic Fairness

Definition 1: Demographic Parity (DP) (Barocas and Selbst 2016) - Let f be a function $f : X \to \hat{Y}, \hat{Y} = \{0, 1\}$ for binary classification, and let S, be a sensitive attribute, the function f satisfies DP if:

$$P[f(x) = 1 \mid x \in S] = P[f(x) = 1], \forall s \in S \quad (2)$$

where x denotes an instance of X and $P[.]$ denotes the probability of an instance.

Definition 2: Equal Opportunity or True Positive Rate Parity (TPR) for a binary prediction $\hat{Y}$ and a member subgroup A, is satisfied if:

$$P(\hat{Y} = 1 \mid A \in \text{SG}_i, Y = 1) = P(\hat{Y} = 1 \mid A \in \text{SG}_j, Y = 1) \quad \forall i, j \in |\text{SG}|, i \neq j \quad (3)$$

2.2 Intersectional Fairness

For defining intersectional fairness in this work, we follow (Ghosh, Genuit, and Reagan 2021).

The worst-case parity (WP) of measuring intersectional DP can be framed as follows (Ghosh, Genuit, and Reagan 2021):

$$\text{WP(DP)} = \frac{\min\{P(\hat{Y} \mid A \in \text{SG}i), \forall i \in |\text{SG}|\}}{\max\{P(\hat{Y} \mid A \in \text{SG}j), \forall j \in |\text{SG}|\}}, \quad (4)$$

where $i \neq j$. Similarly, the worst-case parity of intersectional TPR can be defined as:

$$\text{WP(TPR)} = \frac{\min\{P(\hat{Y} = 1 \mid A \in \text{SG}_i, Y = 1)}{\max\{P(\hat{Y} = 1 \mid A \in \text{SG}_j, Y = 1)}, \quad (5)$$

where $\forall i \in |\text{SG}|, \forall j \in |\text{SG}|$, and $i \neq j$. For a definition of algorithmic fairness, $\mathcal{U}(S, \hat{Y})$ (DP or TPR) on the predictions $\hat{Y}$, the WP can be calculated by taking the min-max ratio of values from the given set of subgroups. A value closer to 1 guarantees fair outcomes and we follow the 80% rule.

***Remark**: We recommend choosing a fairness measure based on the context of the application in healthcare. The DP measure focuses solely on positive outcomes and does not consider ground truth values. However, in situations where outcomes are conditioned on the patient's acute index - for example, if all acute index values are similar - we suggest DP conditioned on acute index levels. Otherwise, we recommend using TPR as it helps in detecting underdiagnosed biases.*

3 Methodology

In this section, we describe the proposed workflow (Supplementary file, Section 1) and method in detail (Figure 2).

3.1 Learning Unified Text Representations from Multimodal Data

The first step in our proposed work is to gather information from multimodal sources. To capture the information into a unified feature representation, we extend the approach described in (Lee et al. 2024), where textual representations of tabular EHR features have been learned. However, it overlooked the heterogeneous multimodal nature of the EHR data, which we address by extending the concept of textual representation learning to the remaining modalities.

We perform pre-processing to remove noises from the clinical texts to obtain EHR_notes. For generating radiology reports from the chest x-rays, we follow (Tanida et al. 2023). Given a chest X-ray, we run it through an object detection model to extract 29 regions of interest and run a binary image classifier to find any abnormalities. A Language model has been employed for creating the abnormality descriptions followed by a post-processing step to create EHR_Xray. For clinical events, we adopt a filtering mechanism to remove repeated events which are subsequently described in textual form, EHR_events. For laboratory results (with time stamps) we follow the box-plot outlier detection (Rousseeuw and Hubert 2018) to locate abnormal portions which are then textualized as EHR_lab. Finally, the tabular representations are textualized $(EHR_structured)$ using the method described in (Lee et al. 2024).

The unified multimodal feature representation learning can be written as:

$$EHR_\text{unified}(X_i) = EHR_\text{text}(X_i) \quad \forall X_i, i \in \{1, 2, \ldots, n\} \quad (6)$$

Here, $EHR_\text{unified}(X_i)$ denotes the unified textual embedding for a patient X_i from multimodal sources and $EHR_$text denotes the resultant textual representation.

3.2 Embedding and Prediction

One could argue that with the emergence of Large Language Models like GPT, training a model from scratch using embeddings is no longer necessary. However, this argument has been refuted in (Lee et al. 2024). To generate embeddings for the unified EHR concepts in the $EHR_$unified textual representation and enable the downstream models to learn interoperable and cross-modal features, we employ clinical LMs: Med-BERT, ClinicalBERT, and Bio-clinicalBERT, which centers on the BERT framework containing contextualized embeddings of the EHR dataset of $28,490,650$ patients. Note that, the unified representation now contains information from all the sources unlike (Lee et al. 2024). During the training stage, tokenized data from $EHR_$unified is fed into the clinical LMs. These encoders produce embeddings that capture information from multimodal sources of a patient's EHR in rich and high-dimensional vector representations. We keep the encoders frozen at this stage and update the weights of the subsequent prediction models dedicated to tasks. The mathematical modeling can be represented as :

$$\vec{E} = \text{Clinical LMs}(c_i), \quad (7)$$

where, $\forall c_i \in Tokenizer(EHR_\text{unified})$. Once we generate the embeddings, the final output is sent to the subsequent classifiers for predictions. We train a self-attention neural network classifier to align the cross-modal features followed by fully connected layers as this combination will enhance the predictive capabilities of the model (Lee et al. 2024). This can be represented as :

$$\vec{E}_{\text{attention}} = \text{SelfAttention}(\vec{E}); \quad (8)$$

$$\vec{E}_{fc} = \text{ReLU}(FC(\vec{E}_{\text{attention}})) \quad (9)$$

The classifier output can then be modeled:

$$\hat{Y} = \min_{\text{loss_}task}(\text{Classifier}(\vec{E}_{fc})), \quad (10)$$

where loss_$task$ is the loss tailored for the individual binary predictions. After obtaining the task predictions, an extensive bias analysis has been done to uncover potential biases.

3.3 Subgroup-specific Discrimination Aware Ensembling (SDAE)

The final step is to design bias mitigation strategies considering intersectionality and multimodality. In this work, we consider post-process bias mitigation for two main reasons:

1. Data preprocessing approaches (Kamiran and Calders 2012) such as Reweighting and disparate impact remover (Feldman et al. 2015) will not be feasible in multimodality as data sources are dispersed, making the identification of data-label pairs impractical.
2. In-processing methods such as adversarial debiasing (Zhang, Lemoine, and Mitchell 2018) often fall short of completely optimizing an adversary (Zhu et al. 2021) and their effect on intersectional subgroups is an ongoing research area, which we leave for future study.

Based on the above observations we consider methods that enforce fairness through post-processing. Our goal is to learn

$\mathcal{U}(\text{SG}, \hat{Y})$- enforced classifier models for the downstream binary and multitask prediction tasks trained on the multimodal patient records, where SG is the set of intersectional subgroups as detailed in Section 2.

Definition 3: Bias mitigation procedure - Let $\mathcal{M}$ be the base classifier, $\hat{Y}$ be the predictions of $\mathcal{M}$, then a bias mitigation procedure $\mathcal{P}$ is defined as a function that can be applied on the true predictions $\hat{Y}$ of $\mathcal{M}$ so that the resultant derived predictions $\hat{Z}$ of $\mathcal{M}$ is $\mathcal{U}(A, \hat{Z}) - fair$, for all subgroup members,$A \in$ SG.

Training: The proposed bias mitigation procedure $\mathcal{P}$ involves a post-processing approach through a series of steps to achieve derived predictions $\hat{Z}$. For this, we rely on bootstrap ensembling (Breiman 1996) and extend it to intersectional and multimodal settings. For mitigation, we build the procedure $\mathcal{P}$ on the unified feature representation EHR_unified and build multimodal models $\mathcal{M}^M$ corresponding to each split, EHR_unified^M, where M is the number of total models/dataset splits for subgroup pairs in SG. For example, if SG={(i).(Female, White), (ii).(Female, Black), (iii).(Male, White), (iv).(Male, Black)}, then the split will be based on the pairs in SG, such as $\{EHR\text{_unified}^M\}$ = $\{EHR\text{_unified}^{i-ii}$, $EHR\text{_unified}^{i-iii}$, $EHR\text{_unified}^{i-iv}$, $EHR\text{_unified}^{ii-iii}$, $EHR\text{_unified}^{ii-iv}$, $EHR\text{_unified}^{iii-iv}\}$, sum up to 6 splits enclosing multimodal subgroup-specific information. These splits are then embedded individually using Med-BERT and subgroup-specific classifiers are trained on the embeddings for predictions. The main idea behind this technique is that each classifier will learn the subgroup-specific features and contribute to the final ensembling stage. The models $\mathcal{M}^{i-j}$, where i, j denotes that the model is trained on the split $EHR\text{_unified}^{i-j}$.

Prediction: In prediction, each instance in the test/validation split is processed by models that correspond to the subgroup membership of that instance and the base multimodal classifier. We then form a Discrimination-Aware Ensemble (Kamiran, Karim, and Zhang 2012) by exploiting the disagreement region of all the subgroup-specific classifiers. Relying solely on voting to calculate the score may result in many instances having identical score values making it challenging to decide which instances should be selected for class changes based on their scores, especially when the number of subgroups is small. To resolve this issue, we perform the following: (i) if all voters are in consensus, we assign the corresponding class label; otherwise, we compute a score η for each instance X_i as a linear combination of the classifier's probabilities $P^{\mathcal{M}^{i-j}}$ and votes V^{i-j}:

$$\begin{aligned} \text{If agreement,} \quad & Z_i = \bar{V}^{i-j}; \\ \text{Else,} \quad & \eta_i = h \cdot \bar{V}^{i-j} + (1-h) \cdot \bar{P}^{\mathcal{M}^{i-j}}; \\ & Z_i = \begin{cases} 1 & \text{if } \eta_i > \tau_{SG}, \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{11}$$

where $\bar{V}^{i-j}$ is the winner (if not in consensus, it will be the ratio of the majority), X_i carries a subgroup membership with models $\mathcal{M}^{i-j}$, $\bar{P}^{\mathcal{M}^{i-j}}$ is the average of the classifier probabilities, Z_i is the derived prediction, and τ_{SG} (0.5 in our case) is a hyperparameter, that can be set to balance the fairness-accuracy trade-off for each of the subgroups. The parameter h is introduced to prevent the predicted probabilities from overshadowing the $\bar{V}^{i-j}$ selection and can be set according to the number of models involved in the voting process as, $\frac{|M^{i-j}|-1}{|M^{i-j}|}$. Note that, aggregating predictions, when all the models are in consensus will not have any significance on fairness. We exploit the disagreement region of individual predictors and use our ensembling strategy to improve fairness.

Modality	**Med BERT**	**Clinical BERT**	**Bio-ClinicalBERT**
Arrival+ Triage+ Medrecon+	0.920	0.913	0.913
Vitals	0.734	0.744	0.744
Codes+ Pyxis	0.740	0.738	0.736
Xray	0.757	0.923	0.919
Multimodal(+events)	**0.925**	**0.741**	**0.913**

Table 1: Multimodality analysis using F1 score and various BERT embeddings in MIMIC-Eye ED Disposition.

4 Experiments

4.1 Downstream Tasks

The MIMIC-Eye dataset (Hsieh et al. 2023)[4] integrates various datasets associated with the Medical Information Mart for Intensive Care (MIMIC), offering a comprehensive collection of patient information. It includes medical images such as chest X-rays (available in both MIMIC CXR and MIMIC JPG formats), clinical data from the MIMIC IV Emergency Department (ED), detailed patient hospital journey records from MIMIC IV, and eye-tracking data capturing gaze information and pupil dilations. Detailed statistics of the dataset is given in the supplementary file, Section 2. We consider both MIMIC-Eye and MIMIC-IV ED for our experiments.

Benchmarking tasks: In this study, we explore the downstream tasks to analyze the predictive performance. Specifically, we perform the binary classification and multitask classification (Lee et al. 2024).

- ED Disposition - This involves predicting the post-visit destination of patients following their ED visit. This is a binary classification problem where label 1 denotes patients who were admitted to ED and 0 denotes patients who were discharged home.
- ED Decompensation - It is a multitask binary classification and is used to predict three ED tasks simultaneously. The first task predicts the patient's discharge location (1:home, 0:not home). The second task predicts the need for an ICU. The third predicts patient mortality, specifically whether the patient will die during their hospital stay.

[4]https://physionet.org/content/mimic-eye-multimodal-datasets/1.0.0/

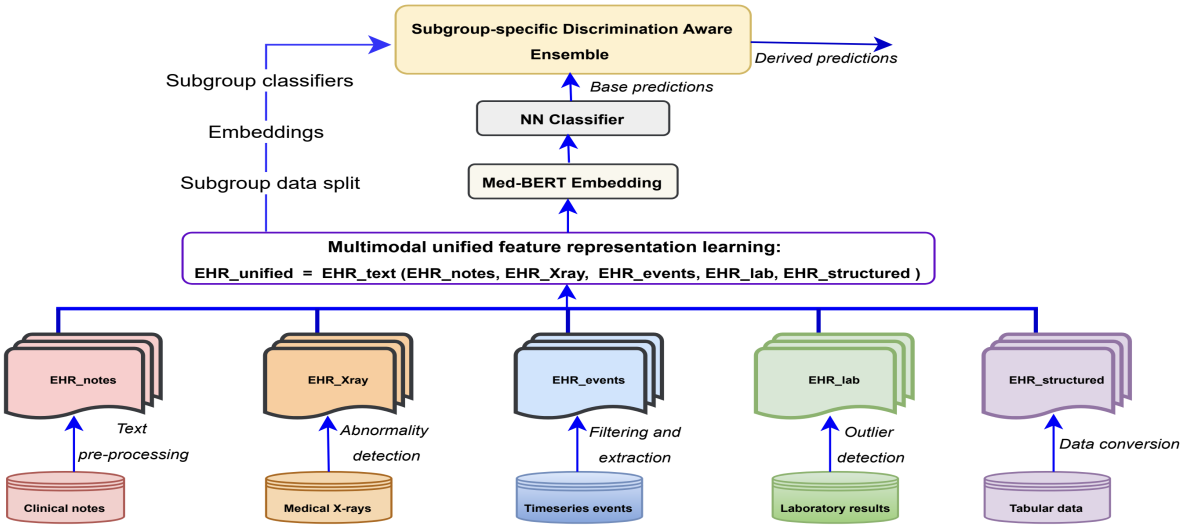


Figure 2: The proposed architecture

We use F1 (Hicks et al. 2022) to compare the performances of these tasks. The experimental settings are given in the supplementary file, Section 3.

Multimodality selection: We analyze the model's performance change when supplemented with modality across embeddings. The underlying idea is that a well-designed model trained from multiple sources should not exhibit performance degradation (Chen et al. 2024) [5] when some information is missing at inference. Analysis of individual performances of modalities against multimodal models can better assist in developing more robust models (Chen et al. 2024). Following this observation, we train the downstream prediction models on an increasing subset of modalities using multiple clinical LMs. We initiate the process with a multimodal model trained only on categorical and numeric attributes, then incrementally add modalities from diagnoses, medications, orders, lab results, radiology reports, etc. Table 1 summarizes our analysis and the performance of multimodal models in the ED disposition tasks using MedBERT, clinicalBERT, and Bio-clinicalBERT embeddings. We provide the results of multitask learning in the supplementary file (Section 6). Based on the analysis, one can choose MedBERT over other embeddings as it better captures the multimodal interoperability contributing to an F1 score of 0.925 compared to other clinical LMs. In this work, we perform analysis using MedBERT and leave the choice of embedding to be determined by the multimodal data distribution and performance evaluation. Prioritizing fairness without considering performance will cause more adverse effects in healthcare and therefore a multi-criteria approach in decision-making is necessary (Baltussen and Niessen 2006).

4.2 Intersectional Bias Analysis and Mitigation

Bias analysis: We perform a comprehensive analysis concerning multiple sensitive attributes, focusing on *race*, *gender*, and their intersectional subgroup pairs with algorithmic fairness metric, $\mathcal{U}(A, \hat{Y}) \forall A \in SG$, as detailed in Section 2.

Table 2 highlights limitations of how examining fairness metrics for a single sensitive attribute can obscure inequalities when intersectional subgroups are involved. The base classifier, $\mathcal{M}^{base}$ shows no significant disparity when evaluating the DP, TPR, and WP solely for gender. Conducting a more nuanced intersectional analysis on race and gender reveals that the disparity (DP) is even more pronounced in Asian Females than in Asian Males in ED Disposition.

Bias mitigation: Here, we compare our SDAE approach against state-of-the-art Reject Option Classification (ROC) (Kamiran, Karim, and Zhang 2012). Note that the ensemble, MAAT (Chen et al. 2022) relies on output probabilities which limits the applicability of post-processing methods. We compare SDAE against (i) the Base classifier with ROC (gender), (ii) the Base classifier with ROC (race), and (iii) the base classifier with ROC (gender+race). The results are shown in Tables 2 and 3 respectively for MIMIC-Eye and the biased sampling from MIMIC-IV ED. The main findings are:

1. **Existing ROC approach injects additional biases at the intersections:** When performed ROC based only on gender, it injects additional biases into the race subgroups as the method tends to assign favorable outcomes to the gender and unfavorable outcomes to the subgroups of race and gender as both consist of privileged and unprivileged groups. This shows that when the intersectional subgroups are correlated, ROC on one sensitive attribute will have negative impacts on the intersectional groups (this holds for ROC on race as well).

2. **Existing methods suffer from the problems of leveling down while SDAE improves the performance of minority:** When enforcing DP using ROC in MIMIC-eye, the value for the racial minority, *Asian*, drops from 0.575 to 0.468. In contrast, our SDAE method improves it to 0.601, resulting in a positive movement for the minority group.

[5] The ME-BEC dataset is not publicly available

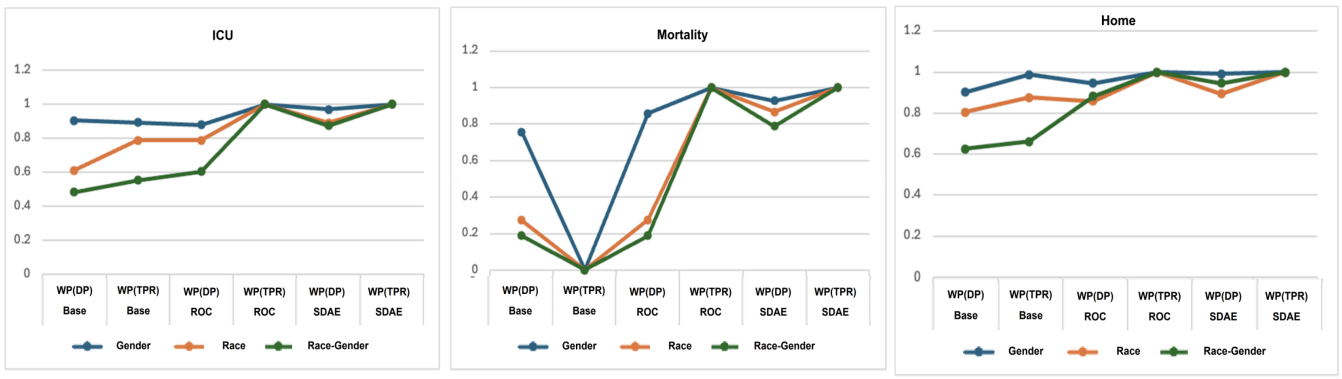


Figure 3: Comparison of WP in multitask setting using MIMIC-Eye

In all cases from both MIMIC-Eye and MIMIC-IV ED datasets, our SDAE method enhances the performance of minority groups. ROC methods achieve fairness by reducing both majority and minority values, but this approach leads to degradation in minority performance which is precisely for whom the mitigation strategies are designed.

3. **The existing ROC with race+gender has a null effect on intersectionality:** This proves the inability of single attribute settings to intersectionality.

4. **The metric worst-case parity (WP) does not reflect the overall fairness reduction:** For the gender group in the MIMIC-Eye disposition task, there is a drop in DP values (.702 to .610) of ROC- models despite achieving a WP of 1. This shows that WP accounts only for the ratio, not the group-wise drop.

Note: The TPR for Female-Asian has been improved with SDAE, but not the DP as the negative probabilities lie close to 0. We recommend choosing the threshold τ_{SG} (Equation 11) to force them to be in a positive class conditioned on an acute index of patients, as a lower-priority acute index does not demand an ED admission. If one wants them to be in a positive class, the threshold τ_{SG} can be set to a very low value for the subgroups, which we don't recommend as DP only considers positive predictions, not the ground truth. The overall WP for multitask comparison in MIMIC-Eye is given in Figure 3 with results for MIMIC -IV ED in the supplementary file, Section 5.

Significance of SDAE: From Tables 2 and 3, and Figures 3 and 4 it is evident that our proposed SDAE model demonstrates consistent improvement in fairness (DP, TPR, and WP) across all tasks and datasets, with negligible performance degradation (Figure 4). The reason can be attributed to the unified text representation learning from modalities coupled with the joint advantages of ensembling and post-processing. The sub-group specific multimodal models are framed to capture the interoperability of features and exploit the disagreement. Unlike traditional ensembling, SDAE customizes the disagreement region to prevent identical scores across multiple subgroups, which effectively improves the fairness values of minorities.

Group	$\mathcal{M}^{base}$		$\mathcal{M}^{ROC}$		$\mathcal{M}^{SDAE}$	
	DP	TPR	DP	TPR	DP	TPR
Male(M)	0.702	0.997	0.610^{-}	1.000	0.710	0.998
Female(F)	0.703	0.998	0.610^{-}	1.000	0.704	0.996
WP	0.998	0.998	**1.000**	**1.000**	0.991	0.997
White	0.708	0.997	0.611^{-}	1.000	0.702	0.998
Black	0.707	0.996	0.622^{-}	1.00	0.688	0.998
Asian	0.575	1.000	0.468^{-}	1.000	0.601	0.998
WP	0.812	0.996	0.752^{-}	**1.000**	**0.856**	**1.000**
White-M	0.697	0.995	0.602^{-}	1.000	0.687	1.000
White-F	0.719	0.998	0.620^{-}	1.000	0.709	1.000
Black-M	0.720	0.996	0.631^{-}	1.000	0.717	0.996
Black-F	0.696	0.997	0.615^{-}	1.000	0.688	0.996
Asian-M	0.638	1.000	0.511^{-}	1.000	0.638	1.000
Asian-F	0.511	1.000	0.426^{-}	1.000	0.510	1.000
WP	0.709	0.995	0.675^{-}	**1.000**	**0.711**	0.996

Table 2: Fairness comparison on MIMIC-Eye using the base classifier ($\mathcal{M}^{base}$), ROC models ($\mathcal{M}^{ROC}$), and the SDAE ($\mathcal{M}^{SDAE}$) in Disposition. We enforced ROC on gender, race, and gender+race, and obtained identical results for all three ROC models, which we denote as $\mathcal{M}^{ROC}$ (best results are bolded) and a $'-'$ indicates the values are reduced from the base model (more than 5% reduction) though WP is improved. Here, DP is calculated based on the positive prediction of each subgroups and TPR is the true positive rate.

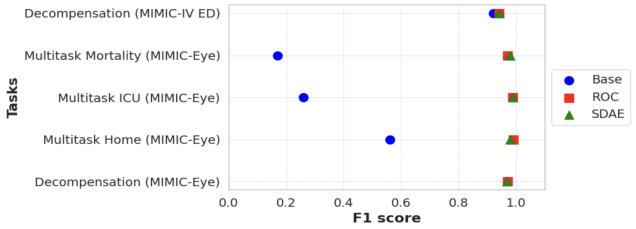


Figure 4: Comparison of F1 scores

Group	$\mathcal{M}^{base}$		$\mathcal{M}^{ROC}$		$\mathcal{M}^{SDAE}$	
	DP	TPR	DP	TPR	DP	TPR
Male(M)	0.399	0.960	0.370^{-}	1.000	0.398	0.991
Female(F)	0.397	0.960	0.366^{-}	1.000	0.396	0.992
WP	0.994	1.000	0.989	1.000	**0.994**	0.998
White	0.420	0.969	0.397^{-}	1.000	0.409	0.990
Non-White	0.339	0.928	0.289^{-}	1.000	0.363	1.000
WP	0.807	0.957	0.727	**1.000**	**0.887**	**0.990**
White-M	0.420	0.969	0.398^{-}	1.000	0.409	0.990
White-F	0.421	0.969	0.397^{-}	1.000	0.410	0.989
NonWhite-M	0.345	0.925	0.288^{-}	1.000	0.367	1.000
NonWhite-F	0.341	0.932	0.292^{-}	1.000	0.361	1.000
WP	0.809	0.954	0.723	**1.000**	**0.880**	0.989

Table 3: Fairness comparison between the base classifier, ROC models, and the proposed SDAE in ED Disposition for a biased sampling in MIMIC-IV-ED dataset. The $'-'$ indicates group wise value reduction (more than 5 percent) and the proposed SDAE achieves DP and TPR without groupwise reduction.

5 Related Works

There have been numerous pieces of research on biases in various clinical data sources and tasks (Pivovarov et al. 2014; Boyd et al. 2023; Banerjee et al. 2023; Tripathi et al. 2023; Arias-Garzón et al. 2023). These studies have overlooked the biases in multimodal data sources with varying intensities and patterns. Recent works have shown efforts in identifying intersectional biases in clinical decision tasks (Ogungbe, Mitra, and Roberts 2019; Okoro, Hillman, and Cernasev 2022). However, to the best of our knowledge, there is no research in identifying and mitigating intersectional biases in tasks utilizing extensive multimodal data from all the sources, especially in the medical domain. Current intersectional bias mitigation strategies (Foulds et al. 2020; Chen et al. 2022) cannot be extended to multimodality as these are tailored for their specific in-process/pre-process algorithms, limiting their flexibility to multimodal models.

6 Conclusions and Future Works

Our work introduced a novel framework, SDAE, designed for bias mitigation involving intersectional subgroups, utilizing multimodal information from clinical databases. We have shown that learning a unified feature representation in the form of text from multimodal data sources improves the overall performance of the model. Additionally, the modality-wise performance metrics illustrate that multimodal models positively contribute to the overall prediction performance in binary and multitask classifications. The bias analysis focused on the intersectional subgroup level revealed the inability of traditional single-attribute-centric approaches to capture the biases at those intersections. To account for this, we proposed a Subgroup-specific Discrimination-Aware Ensemble (SDAE), a post-processing method customized at the intersection level. The extensive intersectional fairness analysis on binary and multitask learning illustrated that the proposed SDAE improves the Worst case Demographic Parity of minorities when evaluated on single attribute and intersectional settings, without reduction in group-wise DP. This is in contrast to conventional methods, which have shown improvement in single attribute scenarios while sacrificing overall fairness values. Also, the Worst case True Positive Rate is reasonably maintained across the subgroups and achieved reasonable performance compared to Reject Option Classification (ROC) methods. Apart from the intersectional bias analysis, modality utilization, and bias mitigation, we benchmark the models and tasks presented in this work highlighting the challenges associated with diverse, complex, and data-rich modalities. We anticipate that future bias mitigation applied to MIMIC-Eye and MIMIC-IV ED will leverage our work, to encompass more tasks and analysis across patient visits.

The future directions for this study will be to extend our SDAE framework to encompass mitigation using the entire pipeline including pre-processing and in-processing methods and framing a metric for measuring the intersectional biases reflecting the overall fairness change across the sub-groups.

Limitations

Despite the numerous advantages outlined in this work in terms of intersectional fairness and multimodal learning without performance degradation, there are limitations to our work. Our Subgroup-specific Discrimination Aware Ensembling (SDAE) will be computationally expensive if the sensitive attributes have high cardinality, resulting in a high number of intersections (in our empirical analysis we have used subgroup pairs up to 9). Also, our SDAE is designed as a post-process mitigation, and its adaptability to in-process and pre-process strategies has not been explored here. It is worth mentioning that MIMIC-Eye being a combination of MIMIC-IV and MIMIC-CXR has only 15627 stay IDs. We have overcome the limitations of small-scale data learning using clinical Language Models and showed that the performance has been reasonably maintained. But, in situations where pre-trained models are not available for applications, the multimodality information extraction will be cumbersome.

Ethics Statement

The performance and bias metrics used in this work are taken from the literature and not from any clinical trials or real-world scenarios. Therefore before using our work in real-world settings, it is advisable to get recommendations from healthcare experts, ethicists, and policy makers. Another important point is that the major sources of biases in this work depend on the distribution of demographic data, model selection, and modality selection. We, therefore, recommend conducting an extensive bias analysis on any real-world dataset before generalizing our method. Finally, we advocate for analyzing group-wise performance drops in fair interventions despite satisfying the overall fairness criteria as interventions can be explicitly fair. This causes performance drops in minority, for whom we aim for fairness enforcement.

References

Alabdulmohsin, I.; Wang, X.; Steiner, A.; Goyal, P.; D'Amour, A.; and Zhai, X. 2024. CLIP the Bias: How Useful is Balancing Data in Multimodal Learning? *arXiv preprint arXiv:2403.04547*.

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Amal, S.; Safarnejad, L.; Omiye, J. A.; Ghanzouri, I.; Cabot, J. H.; and Ross, E. G. 2022. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in cardiovascular medicine*, 9: 840262.

Arias-Garzón, D.; Tabares-Soto, R.; Bernal-Salcedo, J.; and Ruz, G. A. 2023. Biases associated with database structure for COVID-19 detection in X-ray images. *Scientific reports*, 13(1): 3477.

Baltussen, R.; and Niessen, L. 2006. Priority setting of health interventions: the need for multi-criteria decision analysis. *Cost effectiveness and resource allocation*, 4: 1–9.

Banerjee, I.; Bhattacharjee, K.; Burns, J. L.; Trivedi, H.; Purkayastha, S.; Seyyed-Kalantari, L.; Patel, B. N.; Shiradkar, R.; and Gichoya, J. 2023. "Shortcuts" causing bias in radiology artificial intelligence: causes, evaluation and mitigation. *Journal of the American College of Radiology*.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104: 671.

Boyd, A. D.; Gonzalez-Guarda, R.; Lawrence, K.; Patil, C. L.; Ezenwa, M. O.; O'Brien, E. C.; Paek, H.; Braciszewski, J. M.; Adeyemi, O.; Cuthel, A. M.; et al. 2023. Equity and bias in electronic health records data. *Contemporary Clinical Trials*, 130: 107238.

Breiman, L. 1996. Bagging predictors. *Machine learning*, 24: 123–140.

Chen, E.; Kansal, A.; Chen, J.; Jin, B. T.; Reisler, J.; Kim, D. E.; and Rajpurkar, P. 2024. Multimodal Clinical Benchmark for Emergency Care (MC-BEC): A Comprehensive Benchmark for Evaluating Foundation Models in Emergency Medicine. *Advances in Neural Information Processing Systems*, 36.

Chen, Z.; Zhang, J. M.; Sarro, F.; and Harman, M. 2022. MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *Proceedings of the 30th ACM joint european software engineering conference and symposium on the foundations of software engineering*, 1122–1134.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, 424–432. SIAM.

Ghosh, A.; Genuit, L.; and Reagan, M. 2021. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, 22–34. PMLR.

Hicks, S. A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M. A.; Halvorsen, P.; and Parasa, S. 2022. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1): 5979.

Hsieh, C.; Ouyang, C.; Nascimento, J. C.; Pereira, J.; Jorge, J.; and Moreira, C. 2023. Mimic-eye: Integrating mimic datasets with reflacx and eye gaze for multimodal deep learning applications. *PhysioNet (version 1.0. 0)*.

Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L. A.; and Mark, R. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 49–55.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Kamiran, F.; Karim, A.; and Zhang, X. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, 924–929. IEEE.

Kline, A.; Wang, H.; Li, Y.; Dennis, S.; Hutch, M.; Xu, Z.; Wang, F.; Cheng, F.; and Luo, Y. 2022. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1): 171.

Lee, S. A.; Jain, S.; Chen, A.; Biswas, A.; Fang, J.; Rudas, A.; and Chiang, J. N. 2024. Multimodal Clinical Pseudo-notes for Emergency Department Prediction Tasks using Multiple Embedding Model for EHR (MEME). *arXiv preprint arXiv:2402.00160*.

Niu, S.; Ma, J.; Bai, L.; Wang, Z.; Guo, L.; and Yang, X. 2024. EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Information Fusion*, 102: 102069.

Ogungbe, O.; Mitra, A. K.; and Roberts, J. K. 2019. A systematic review of implicit bias in health care: A call for intersectionality. *IMC Journal of Medical Science*, 13(1): 5–21.

Okoro, O.; Hillman, L.; and Cernasev, A. 2022. Intersectional invisibility experiences of low-income African-American women in healthcare encounters. *Ethnicity & Health*, 27(6): 1290–1309.

Pang, Z.; Wang, X.; Wang, X.; Qi, J.; Zhao, Z.; Gao, Y.; Yang, Y.; and Yang, P. 2021. A multi-modal data platform for diagnosis and prediction of Alzheimer's disease using machine learning methods. *Mobile Networks and Applications*, 26(6): 2341–2352.

Pivovarov, R.; Albers, D. J.; Sepulveda, J. L.; and Elhadad, N. 2014. Identifying and mitigating biases in EHR laboratory tests. *Journal of biomedical informatics*, 51: 24–34.

Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; and Zhi, D. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1): 86.

Rousseeuw, P. J.; and Hubert, M. 2018. Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2): e1236.

Rouzrokh, P.; Khosravi, B.; Faghani, S.; Moassefi, M.; Vera Garcia, D. V.; Singh, Y.; Zhang, K.; Conte, G. M.; and Erickson, B. J. 2022. Mitigating bias in radiology machine learning: 1. Data handling. *Radiology: Artificial Intelligence*, 4(5): e210290.

Seymour, T.; Frantsvog, D.; and Graeber, T. 2012. Electronic health records (EHR). *American Journal of Health Sciences*, 3(3): 201.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *CVPR*.

Tripathi, S.; Gabriel, K.; Dheer, S.; Parajuli, A.; Augustin, A. I.; Elahi, A.; Awan, O.; and Dako, F. 2023. Understanding biases and disparities in radiology AI datasets: a review. *Journal of the American College of Radiology*, 20(9): 836–841.

Turan, J. M.; Elafros, M. A.; Logie, C. H.; Banik, S.; Turan, B.; Crockett, K. B.; Pescosolido, B.; and Murray, S. M. 2019. Challenges and opportunities in examining and addressing intersectional stigma and health. *BMC medicine*, 17: 1–15.

Yan, S.; Huang, D.; and Soleymani, M. 2020. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 361–369.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhu, W.; Zheng, H.; Liao, H.; Li, W.; and Luo, J. 2021. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15002–15012.

7 Reproducibility Checklist

This paper:

1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes)
2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
3. Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)
4. Does this paper make theoretical contributions? (no)
5. A motivation is given for why the experiments are conducted on the selected datasets (yes)
6. All novel datasets introduced in this paper are included in a data appendix. (yes/links are given)
7. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (NA)
8. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
9. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (partial/have to complete training)
10. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (yes)
11. Does this paper include computational experiments? (yes)

    If yes, please complete the list below.
12. Any code required for pre-processing data is included in the appendix. (yes).
13. All source code required for conducting and analyzing the experiments is included in a code appendix. (yes)
14. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
15. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
16. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
17. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
18. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
19. This paper states the number of algorithm runs used to compute each reported result. (yes)
20. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
21. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (yes)
22. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
23. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (yes)