UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

A comparative study of fairness methods for clinical predictions using the MIMIC-IV database

Author:
Iris VUKOVIC

Supervisor:
Dr. Laura IGUAL MUÑOZ

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

January 19, 2026

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

A comparative study of fairness methods for clinical predictions using the MIMIC-IV database

by Iris VUKOVIC

Fairness methods are an increasingly important aspect of responsible implementations of machine learning models. As machine learning becomes more intertwined in clinical settings, it is necessary that bias mitigation is accounted for, but performance maintenance remains a challenge. Fairness-aware interpretable modeling (FAIM) [1] is an in-processing fairness method that avoids extreme performance degradation while improving fairness and maintaining interpretability. In this study, the method is stress-tested by changing the original prediction task, hospital admission prediction after emergency department (ED) stay, to the distinct clinical task of predicting necessity of invasive medical ventilation (IMV) for patients in the intensive care unit (ICU) using electronic health record (EHR) data from the recently released MIMIC-IV database. A comparison with the baseline logistic regression model and other state-of-the-art fairness methods is presented and, although bias amongst intersectional demographic subgroups was not completely mitigated with FAIM, there was clear improvement compared to the baseline and also compared to other traditional fairness methods.

Acknowledgements

Thank you to Laura, my supervisor, for your guidance.

Thank you to my friends and family for always supporting me with love and laughter.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As it is collected from a world where existing racism, sexism, and classism are rampant, most real-world data is embedded with societal biases. Machine learning models trained with these datasets are bound to learn and amplify inequities for individuals disproportionately impacted by systemic biases. As the use of machine learning expands into increasingly sensitive domains, it is essential that fairness methods expand with them. However, in the case of most state-of-the-art fairness methods, there is a steep tradeoff between model performance and bias mitigation, where an improvement in one results in a deterioration in the other. Regardless of this fact, it is crucial that rigorous fairness assessment is conducted before the deployment of any model, arguably in every field in which machine learning is being implemented, but especially in healthcare due to the high-risk nature of clinical decision-making.

Interpretability, referring to transparency in model architecture and decision-making criteria, is a notion that differs from fairness. Although both contribute to advancing ethically-informed machine learning, they are not as complementary as they seem. That is, their relationship is complex and implementing one method can diminish the effects of the other. Nonetheless, both are strong tools for bias mitigation. Without the implementation of fairness and interpretability methods, the widespread adoption of machine learning models in fields like healthcare is largely infeasible.

Liu et al. propose fairness-aware interpretable modeling (FAIM) [1] for trustworthy machine learning in healthcare: a fairness method that successfully mitigates demographic bias while maintaining model accuracy and transparency, thereby alleviating the decline in performance and interpretability often caused by fairness methods as described above. In this work, we first replicate the paper's results by predicting hospital admission after emergency department (ED) stay and compare FAIM to the baseline logistic regression model. After this, in order to test the robustness of the FAIM method, we change the clinical task to predict invasive mechanical ventilation (IMV) allocation. We choose this clinical task motivated by the findings of Abdelmalek et al. [2]. The authors of this study use data from the MIMIC-IV electronic health record (EHR) dataset to prove that there is a lower rate of invasive ventilation in Asian, Hispanic, and Black patients with respiratory failure than there is in White patients with the same health condition. To explore intersectionality, or the unique discriminatory effects arising from the convergence of social categories, we expand the analysis to examine how race and sex jointly influence outcomes and use FAIM to mitigate the bias that the "fairness-unaware" logistic regression model inevitably has. With FAIM, we are able to demonstrate that there exists a "fairness-aware" nearly-optimal model that maintains both high accuracy and interpretability while mitigating multi-dimensional demographic bias.

The structure of the document is as follows: Chapter 1 outlines the motivation for the study. Chapter 2 introduces the MIMIC-IV database and summarizes relevant data pre-processing steps. Chapter 3 describes the clinical tasks explored and the features included. Chapter 4 defines fairness metrics and details fairness methods, including FAIM. Chapter 5 discusses experimental findings. Finally, Chapter 6 concludes with the limitations of the study and possible future directions. The appendix contains supplementary tables and figures as well as a link to the project's GitHub repository and the bibliography includes sources referenced throughout the paper.

Chapter 2

MIMIC-IV Database and Clinical Tasks

2.1 MIMIC-IV

The MIMIC-IV database is a publicly available collection of electronic health records (EHR), or comprehensive digital accounts of patient medical histories [3]. It contains data from routine clinical procedures as collected from the Beth Israel Deaconess Medical Center located in Boston, Massachusetts in collaboration with the Massachusetts Institute of Technology (MIT). The medical data comes from patients over the age of 18 who had a stay in either the intensive care unit (ICU) or the emergency department (ED) between 2008 and 2022. A set of 18 identifiers, including name and age, as allocated by the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision have been deidentified using custom algorithms developed by researchers at MIT. It is the fourth edition of the MIMIC database. The first edition was published in 2000 with data from just 90 ICU patients [4] and the current edition updated in 2024 stores data from over 65,000 ICU patients and 400,000 ED stays [5]. MIMIC-IV is similar in structure to its predecessors MIMIC-III and MIMIC-II with minor changes including new columns or column names. Also, patient identifiers like `stay_id` are regenerated in MIMIC-IV, so users granted access to multiple versions of MIMIC can't link patients between datasets based on repeated identifiers. The dataset is made available to researchers through the PhysioNet webpage [6], requiring the completion of human research training through CITI Program [7] and the signing of a data use agreement that prohibits de-identifying or sharing the data.

The MIMIC-IV database is divided into six modules, **hosp**, **icu**, **note**, **cxr**, **ed**, and **ecg**, which are further organized into various tables [8] (A.1). In this study, we primarily use data from the **hosp** module, which contains data collected from the hospital-wide EHR, and the **icu** module, which contains data from the clinical information system used within the ICU. For the purpose of replicating the results of the original FAIM paper [1], we use data from the **ed** module, which contains data from patients collected while they are in the ED. We also use the MIMIC-IV **derived** module, which contains transformed and aggregated information from the **hosp** and **icu** modules, such as Severity of Organ Failure Assessment (SOFA) score and Charlson Comorbidity Index (CCI) [9].

2.2 Clinical Tasks

The clinical task analyzed in [1] was the prediction of hospital admission upon ED stay. Hospital admission is a high-stakes prediction task that could be of significant support to clinicians if implemented properly. Early identification allows for earlier provision of hospital beds and medical care for high-risk patients who need it. It also reduces crowding in the ED, which can quickly become very dangerous if not controlled.

To stress-test the FAIM method, we change the prediction task from hospital admission after ED stay to IMV occurrence within the first 48 hours of a patient's stay in the ICU. This time constraint was chosen to mirror real-world decision-making timelines in clinical settings insofar as the model would have to learn whether a patient will need IMV in the next 24 hours from information presented within the first 24 hours. IMV is a form of medical intervention for patients with respiratory failure in which a machine called a ventilator essentially breathes for a patient by delivering oxygen to their airways. In considering patients with respiratory failure, outcomes vary across demographic subgroups [10, 11]. Abdelmalek et al. found that Black, Hispanic, and Asian patients have a higher mortality rate than White patients that can be at least partly explained by differences in IMV allocation [2].

Chapter 3

Data Preprocessing and Features

3.1 Data Preprocessing

In order to validate existing findings, we replicate the FAIM pipeline using the data preprocessing steps as described in [1] following the provided GitHub repositories [12] [13]. Upon changing the clinical prediction task to IMV prediction, we follow similar preprocessing steps as mentioned above and alter some task-specific steps as described below. After being granted access to the database, we query the necessary tables in Google BigQuery from Google Colab and pull the results into pandas DataFrames for downstream analysis. We extract the ventilation status from the *ventilation* table in the **derived** module and define IMV to be the target variable if it was present within the first 48 hours of ICU admission.

For this study, we define a patient with IMV as having either a standard IMV, which consists of a tube going in through the patient's nose or mouth and into the trachea (the tube that conveys air to and from the lungs), or as having a tracheostomy, which involves an incision directly into the neck to attach a tube to the trachea. A patient without IMV either has no breathing assistance or has a form of non-invasive mechanical ventilation, wherein they are provided with breathing assistance through some form of a mask or helmet. We find that the class labels were moderately imbalanced, with a smaller positive (IMV) class (A.3) (A.1).

Each patient is allocated a unique patient identifier (`subject_id`), each ICU stay is allocated a unique stay identifier (`stay_id`), and each hospital admission is allocated a unique admission identifier (`hadm_id`). Therefore, if a patient has multiple ICU stays each with a unique `stay_id`, each of those ICU stays would get the same patient-level information from one `subject_id`. Similarly, a patient can have multiple ICU stays under one `hadm_id`, so all of those ICU stays get the same hospital-level information. We keep only the first ICU stay and eliminate all duplicates in order to prevent data leakage. This way, the model learns to predict whether the patient needs early IMV based on their initial presentation data, instead of learning patterns from subsequent ICU stays.

3.2 Sensitive Features

Race is a social construct guided by supposedly shared physical characteristics and ethnicity is a characterization based on shared cultural background [14]. In the admissions table from the **hosp** module of the MIMIC-IV database, race and ethnicity are recorded in a single column labeled race. Following standards described in [2], we arranged the provided race column into five classes, White, Black, Asian, Hispanic, and Other, and removed any samples in which the patient's race was unknown (A.5).

Gender is a social construct concerning the way one experiences their gender (gender identity) or expresses themselves (gender expression) [15]. Sex relates to one's biological traits. Although we are conditioned to conform to Western ideals about cisnormativity, or the belief that everyone identifies with the sex they are assigned at birth, gender and sex do not always "align" in this way [16]. In the MIMIC-IV database, biological sex is stored as gender and the collected samples refer to the binary indicators, Male and Female.

Intersectionality is the way that social categorizations, like sex and race, interact and create unique biases and discriminatory effects independent of those stemming from individual sensitive attributes.

The most privileged subgroups of the sensitive attributes, Male and White, are marked with an @ symbol in the results tables as the reference groups. Thus, no categorical variable representation is created for them, producing the effect of the White Male subgroup acting as a baseline against which the effects of the other subgroups of sensitive attributes are compared to during FAIM analysis. This means the model learns how being Female, Black, Hispanic, Asian, or Other affects outcomes relative to being Male or White, allowing FAIM to identify and correct disparities relative to the most privileged demographic intersection.

3.3 Rest of Features

As described in their study, Liu et al. train their model on 10 features clinically relevant to hospital admission in addition to the two sensitive variables mentioned above [1]. For our exploration, we keep the same sensitive variables and also include 17 features clinically relevant to the IMV prediction task (A.4). Following MIMIC-IV's de-identification protocol, we reconstruct the de-identified age using `anchor_year` and `anchor_age` from the *patients* table from the **hosp** module by adjusting the difference between ICU admission year and anchor year. We create a dictionary of 8 comorbidities relevant to respiratory failure and their associated ICD-10 (International Classification of Diseases, 10th Revision) codes. Using the `diagnoses_icd` table in the **hosp** module, which records any comorbidities the patient was billed for during their hospital stay, we check each row to see if any comorbidity from our dictionary is present and set binary flags, 1 (present) or 0 (not present), per diagnosis row. Since one unique hospital stay could have been billed for many different comorbidities, and thus have many rows, we collapse rows per unique `hadm_id`. We then query SOFA and CCI from the *derived* table. We gather vital signs, including heart rate and blood pressure, from the *chartevents* table of the **icu** module, convert temperature from Fahrenheit to Celsius in correspondence with the international standard, and remove outliers following range guidelines from FAIM preprocessing steps [12]. Also per FAIM preprocessing, we split the dataset into 70% for training, 10% for validation, and 20% for testing. Missing values in vital signs (temperature, heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure) and SOFA scores are imputed using median strategy for continuous variables [12].

Chapter 4

Fairness

In the context of machine learning, **fairness** is defined as a model not relying on sensitive attributes to make its predictions in a way that discriminates against any individual or group [17, 18]. **Bias** is the the model's predictive performance disparity between subgroups [19].

4.1 Fairness metrics

Fairness can be a difficult metric to quantify due to the assortment of ways that it can be measured which are chosen depending on the context of the task. The FAIM method ranks nearly optimal models based on three fairness metrics for binary classification problems: **equalized odds**, **equal opportunity**, and **balanced error rate (BER) equality** [1] (A.6). The aforementioned metrics are all separation-based, as opposed to independence-based, indicating that they all monitor whether there is equally accurate model performance across subgroups with the protected attribute, in this case race, sex, or their intersections. Independence-based metrics check that the prediction is independent of the protected attribute, or that each subgroup has equal probability of receiving the positive prediction. In the context where ground-truth is available, such as this one, the use of separation-based metrics is favorable since its evaluation of fairness based on error rate among groups is pertinent to model performance [1, 20].

Equalized odds checks that the true positive rate (TPR) and the false positive rate (FPR) are comparable, or that the likelihood of either a true or a false prediction does not deviate much across subgroups [1]. Equal opportunity is a more lenient version of equalized odds that only checks that TPR, or the likelihood of a true positive prediction, is comparable across subgroups. BER equality balances false positive rate (FPR) and false negative rate (FNR) to ensure that misclassification is even across subgroups.

The fairness-aware model is selected by its **Fairness Ranking Index (FRI)**, a measurement developed by Liu et al. that aggregates all of the above fairness metrics into one conclusive score [1]. The calculation is designed so that a higher FRI score indicates a fairer model. Inspired by the geometry of a radar chart, FRI is calculated as the reciprocal of the sum of products of fairness metric pairs, so it captures not only the individual metrics but also their interdependencies (4.1). When two metrics are large, indicating significant bias, their product is even larger, thereby more drastically shrinking the overall FRI.

$$\text{FRI} = \frac{1}{\sum_{i,j} \text{metric}_i \cdot \text{metric}_j + \epsilon}, \quad \text{where } i, j \in \{\text{EOpp, EOdds, BEREq}\} \tag{4.1}$$

Even though FRI only considers the three aforementioned separation-based metrics to select the fairness-aware model, when comparing FAIM with other modern fairness methods we also consider the independence-based metric statistical parity. Statistical parity, or demographic parity, requires an equal selection rate (SR), or rate of positive predictions, in each group, enforcing that the sensitive attribute is independent of the prediction (A.6). However, Liu et al. warn against adding statistical parity to the FRI calculation because of the potential disadvantages of applying it in a clinical setting where forcing equal treatment is illogical when patient needs actually differ. Also, "biological differences and social biases often intertwine," which is to say that sometimes biological differences do impact the presence of medical conditions but can present as biases when looked at from the perspective of statistical parity [1]. This is why we don't utilize it in selecting the final fairness-aware model and consider it cautiously afterwards for fairness metric comparison amongst fairness methods .

4.2 Fairness Methods

To quantify FAIM's efficiency in bias mitigation, we compare it to other state-of-the-art fairness methods. The three techniques for applying fairness methods to machine learning models are as follows (A.8).

Pre-processing Pre-processing adjusts the training dataset before fitting it to the model, with pre-processing methods including reweighing and unawareness. Reweighing assigns weights to each sample in the training dataset based on its sensitive attributes. The weights are calculated by dividing the expected probability, which is the likelihood of an outcome if there was no discrimination in the dataset and predictions were actually independent of sensitive attributes, and the observed probability, which reflects the bias in the dataset [21]. This results in higher weights for underrepresented combinations of attributes, usually minority demographic groups. Another pre-processing method is unawareness, or underblindness, which refers to removing the sensitive attributes from the training dataset all together before learning the model. Pre-processing tends to lower accuracy because it optimizes the model on debiased data and tests it on data with the true, biased distribution.

Post-processing Post-processing alters the outputs of an already-trained model. Equalized odds post-processing forces TPR and FPR to be equal across subgroups. It locates an intersection point where both subgroups have equal TPR and FPR that minimizes the loss in the two-dimensional space of (FPR, TPR) and is reachable by random flipping of model predictions [22]. This method can boost model fairness, but doesn't ensure that the model's new predictions align with real-world likelihoods, reducing transparency because of the randomness introduced [1]. Alternatively, calibrated equalized odds uses group-specific thresholds to optimize accuracy while also balancing TPR and FPR, thereby maintaining prediction calibration and consistent predictions. Reject Option Classifier (ROC) is another post-processing method which first identifies predictions that fall within an uncertain probability range. Then, it flips them for the sake of fairness depending on which fairness metric is being optimized, in this case equal opportunity. If an underprivileged subgroup has a low TPR, some of the model's uncertain negative predictions are flipped to positive ones and vice versa for privileged subgroups with a high TPR.

In-processing In-processing alters the model itself in the pursuit of fairness. Methods include reductions, adversarial learning, and FAIM. Reductions reduces the complexity of the problem of training a fair and accurate classifier into the simpler classification problem of maximizing accuracy on a weighted dataset, hence the name. Using one fairness metric as a constraint, in this case equalized odds, it iteratively re-trains the classifier adjusting weights on each sample based on its subgroup and target outcome label, thus forcing the model to meet the constraint [23]. Adversarial learning for bias mitigation is another in-process fairness method utilizing adversarial debiasing for bias mitigation [19]. It is a dual-network architecture in which the classifier tries to predict the label and the adversary tries to predict the sensitive attributes of the sample. Basically, the classifier is trying to generate predictions from which the adversary is not able to discern the sensitive attributes, intrinsically learning to make predictions that are independent of those attributes. Since it modifies the existing model, in-processing is known to decrease interpretability even when it is successful at mitigating bias.

4.2.1 FAIM

As expressed in its name, FAIM is a fairness method based in fairness-aware interpretable modelling. A "fairness-unaware" logistic regression model is used as a baseline against which to measure FAIM's effectiveness in bias mitigation. Following the approach detailed in [1], FAIM starts off by generating nearly-optimal models using the Rashomon Effect, a phenomenon introduced by statistician Leo Breiman and named after the Japanese film *Rashomon*, which tells a fictional tale about an assortment of differing witness accounts all given for the same murder. In the context of machine learning, it reveals that datasets may admit "many approximately-equally accurate models" [24]. That is, as one approaches the optimal model, one can collect a set of nearly-optimal models that fit the data with a comparable loss, each model relying on the data in a different way and resulting in diverse fairness profiles [1]. For each of the four exclusion scenarios (no exclusion, sex exclusion, race exclusion, and complete exclusion), FAIM first finds the optimal baseline model and then generates a Rashomon set of nearly-optimal models using the ShapleyVIC algorithm [25].

ShapleyVIC serves dual purpose in FAIM, first generating nearly-optimal models and then interpreting these models by calculating their Shapley values. The latter task is further detailed below (Section 4.3). In regards to creating the models, it produces ones that fall within a performance threshold that defaults at allowing up to 5% ($\varepsilon = 0.05$) degradation from optimal area under the curve (AUC) as evaluated on the validation set. Then, it samples from that ε-level set while exploring diverse coefficient combinations. The union of nearly-optimal models from all exclusion scenarios is the Integral Rashomon Set (IRS). Using models from the IRS, fairness metrics and FRI are evaluated on the validation set to allow for the selection of a final fairness-aware model. Once a model is selected, its final performance and fairness status are evaluated on the test set. Since the models output continuous probabilities, those are first converted to binary predictions using Youden's J statistic as the optimal decision threshold. Originally created as a performance metric for medical diagnostic tests, it is a value on the Receiver Operating Characteristic (ROC) curve that maximizes the sum of sensitivity and specificity, or TPR and TNR, balancing the model's ability to identify both positive and negative cases [26]. This ensures that each model is evaluated at its individually optimal operating point rather than using a fixed threshold, most commonly set at 0.5, which may disadvantage models with different calibration properties.

4.3 SHAP

The Shapley value is a game theory based concept often used in economics wherein the contribution of each player is measured to ensure that everyone gets a fair payout based on their participation. In machine learning, the Shapley value measures how much a feature contributes to the final prediction. SHAP is a framework for estimating Shapley values for model interpretability which combines Shapley value theory with various post-hoc interpretation methods [27]. For the purpose of this study, Liu et al. expanded Shapley values into Shapley VIC, or Shapley Variable Importance Cloud [25]. It is another Shapley value based interpretability method that calculates Shapley scores for a number of nearly optimal models and then aggregates the results, thereby capturing uncertainty in Shapley values across different models. In relation to fairness, this interpretability method potentially uncovers bias by illustrating the difference between feature importance rankings when comparing the fairness-unaware model against the fairness-aware one.

Chapter 5

Experimental Results

5.1 Replication Results of Hospitalization Admission Task

Given that a detailed description of this experiment already exists in [1], we only briefly describe our findings here. First, we note that the distribution of the target variable, hospitalization admission after ED stay, is approximately balanced (A.2). Then, we verify that there is indeed bias in the baseline model by comparing fairness metrics (A.6) between the privileged reference group, White Male patients, and the underrepresented groups, Asian, Hispanic, and Black Female patients. We are able to demonstrate that the bias is in fact associated with sensitive attributes sex and race, and is most significant between intersectional demographic subgroups Hispanic Female (TPR, 0.4836) and White Male (TPR, 0.8122) (A.9). Using the FAIM method, we are able to generate a nearly-optimal, fairness-aware model that mitigates demographic bias. Its effectiveness is illustrated by the reduction of fairness metrics across demographic subgroups, namely Hispanic Female (TPR, 0.6236) and White Male (TPR, 0.7612), resulting in a 58.12% decrease in equal opportunity score (A.9). When examining fairness improvements in equal opportunity gap among individual sensitive attributes, FAIM achieved a 55.62% reduction across race subgroups and a 65.71% reduction across sex subgroups (A.11)(A.10). As noted in the original paper [1], our SHAP analysis also showed that the FAIM method produced a model that still valued clinically relevant features without unnecessarily relying on race or sex. In the paper, the authors were able to find a FAIM model that also minimized the importance of features potentially correlated to sensitive attributes, like pain scale. However, our replication model does not minimize the importance of pain scale, and actually increases its feature importance ranking (A.2b).

5.2 Results of Invasive Mechanical Ventilation Task

Before initiating the FAIM sequence, we conduct an exploratory bias analysis on the prediction task to demonstrate that there is demographic bias for IMV allocation. Using the same baseline logistic regression model as implemented in the FAIM workflow, we account for the imbalanced target labels (A.3) by applying balanced class weights and then learn the model on the training dataset. We identify bias in the model by calculating and comparing fairness metrics (A.6) amongst demographic subgroups. We prioritize equal opportunity because false negatives, or missed diagnoses, have severe consequences in clinical settings. There is disparity in the model's performance on intersectional subgroups, most notably between the Black Female subgroup (TPR, 0.6815) and the Asian Male subgroup (TPR, 0.8933), indicating that, according to this model, approximately 32% of Black Female patients who needed IMV wouldn't have gotten it in comparison to 11% of Asian Male patients. The gap between highest and lowest performing subgroups (0.2118) is smaller than the gap found in the original clinical prediction task (0.3286), illustrating the varying levels of bias found in different clinical contexts. Also, this differs from our hypothesis that the most favorable outcome group would be the White Male subgroup (TPR, 0.8564) (5.2). Upon further inspection, we find that the confidence interval for the Asian Male subgroup (CI, 0.813, 0.960) is much wider than that of the White Male subgroup (CI, 0.840, 0.873), indicating more uncertainty for the TPR estimate based on subgroup size for the former (A.3). The Asian Male subgroup takes up only 3.3% of the test set (N: 227) compared to the reference White Male subgroup representing 42.7% (N: 4795). Since small sample sizes can create misleading results, we include the reference subgroup in our comparison for a more complete and trustworthy exploration.

Following the FAIM pipeline, we generate 360 nearly-optimal models and find that the highest ranking model based on FRI excludes sensitive attributes sex and race (A.12). Using SHAP, we find that the baseline model relies on both sex and race, with the former having a higher impact on the model's predictions. The FAIM model excludes both sensitive attributes while retaining clinically relevant features at a high ranking (A.4). FAIM shifts the maximum TPR from Asian Male patients (0.8933) to Asian Female patients (0.8780). Looking at intersectional equal opportunity scores for demographic subgroups, compared to the baseline model (0.2118), the FAIM model (0.1902) decreases the difference between the highest and lowest scoring demographic subgroups by 10.2%, and further improves fairness metrics for sex and race by 43.6% and 19.9% respectively (5.1).

TABLE 5.1: Gap reduction in Equal Opportunity (TPR) across demographic groups for IMV Task

Sensitive Attribute	Baseline Max Gap	FAIM Max Gap	Gap Reduction	Gap Reduction %
Sex	0.0939	0.0530	0.0409	43.6 *
Race	0.1440	0.1154	0.0286	19.9
Sex × Race	0.2118	0.1902	0.0217	10.2

* Indicates statistically significant gap reduction.

Checking for statistical significance in gap reductions, we implement both bootstrapping, wherein we resample the test data with replacement 1,000 times to build confidence intervals around each fairness gap, and permutation testing, wherein we randomly permute model labels 1,000 times to test whether the observed gap reduction could occur by chance under the null hypothesis of no difference between models, and find that only the reduction in gaps between sex subgroups is statistically significant (A.5)(A.7a). In regards to race and intersections, the improvement in bias is small relative to its uncertainty. However, this could very well be due to our relatively small sample size for this clinical task (N: 56,150) (A.3), especially given that in our replication task, which uses a much larger dataset (N: 418,025) (A.2), we find that all gap reductions are statistically significant (A.6)(A.7b).

FAIM is able to maintain accuracy and reduce the equal opportunity gap by selectively improving disadvantaged subgroups and moderating extreme outliers among privileged subgroups (A.15). For example, TPR for Black Female patients increases by 0.94% (0.6815 → 0.6879) while decreasing for White Male patients by 4.7% (0.8564 → 0.8164). Interestingly, even though Asian Female patients are already among the highest-scoring intersectional subgroups (0.8537) in the baseline, already scoring higher than the reference group, they are further boosted by 2.9% (0.878) to outperform the previous best-performer the Asian Male subgroup. However, Asian Female patients take up even less of the test set at 1.7% (N: 184), highlighting the challenges of fairness interpretation with varying intersectional demographic sample sizes. Additionally, the role of the Asian Female subgroup reversed between tasks: in the hospitalization task, they became the minimum-performing group under FAIM's equalized odds metric (FPR: 0.2163), whereas in the IMV task they are the maximum-performing group (TPR: 0.8780). This reversal implies that FAIM's fairness optimization produces task-specific redistributions of performance rather than systematically favoring or disadvantaging particular demographic intersections.

Overall, it seems as though FAIM achieves fairness by decreasing variance rather than by boosting all underprivileged subgroups, reshaping the performance distribution in a way that doesn't always directly benefit the most disadvantaged groups. That is, in the task of IMV prediction, FAIM seems more focused on pulling down higher performers (White Male: -4.7%, Asian Male: -7.5%) than on helping lower-performing subgroups. Though some slightly improved (Black Female: +0.9%, White Female: +0.2%), others actually declined (Hispanic Female: -2.6%, Black Male -2.1%) (A.15). On the other hand, in the original clinical task of hospital admission prediction, FAIM was able to increase the worst-performing subgroup Hispanic Female patients from TPR 0.484 to 0.624, a 29.0% relative improvement and a significantly larger increase for the most underprivileged group than in the IMV task, while the reference subgroup White Male patients had a modest 6.3% decrease in comparison, from 0.812 to 0.761. Additionally, FAIM is able to more drastically reduce the TPR gap in the original task than it is in the IMV prediction task, both among intersectional demographic subgroups (58.2% vs. 10.2%) and among sex (65.71% vs. 43.9%) and race (55.62% vs.19.9%) individually. Briefly disregarding dataset size discrepancy between the two tasks, this could also suggest that the method is more effective when applied to more severely biased contexts in which there is more room for improvement.

Additionally, compared to other state-of-the-art fairness methods, FAIM maintains competitive predictive performance and improves overall fairness metrics compared to the baseline (5.3). Bootstrap significance testing reveals that, despite not optimizing for it, FAIM significantly outperforms the baseline on statistical parity (p=0.024). This suggests that addressing outcome-based fairness metrics, like equal opportunity, can have unintentionally positive effects on selection-based metrics, like statistical parity, without the risks of directly optimizing for them. FAIM also outperforms several fairness methods (Reductions, Equalized Odds, Calibrated Equalized Odds, and Reject Option Classifier) on accuracy equality (p=0.018). While no significant differences were detected for equal opportunity, equalized odds, or BER equality metrics, this pattern suggests FAIM achieves balanced fairness improvements without the extreme trade-offs observed in some methods. For instance, while Reductions achieves better statistical parity (0.114 vs. FAIM 0.146), it simultaneously degrades equal opportunity substantially (0.242 vs. FAIM 0.190) and sacrifices predictive performance (AUC 0.698 vs. FAIM 0.828). Adversarial learning for mitigating bias, or adversarial debiasing, is a promising fairness method which has the best predictive performance (AUC, 0.8476) in comparison with other fairness methods explored. Though it outperforms our FAIM method in specificity (Adversarial learning, 0.7716 vs. FAIM, 0.7253), it has the worst equalized odds score (0.4357) of all methods (5.3) and high variance across metrics. Although FAIM and Unawareness both exclude sensitive variables, FAIM outperforms Unawareness by explicitly measuring fairness during training and optimizing for it, rather than solely relying on the exclusion of sensitive variables to improve fairness metrics. Comparing SHAP scores between them (A.8), two features that display some correlation with sex and race (A.9), age and diastolic blood pressure (DBP) respectively, are ranked higher in feature importance for FAIM than for Unawareness.

There is a notable difference between our study and the original FAIM paper concerning the treatment of features correlated with sensitive attributes, suggesting several possibilities. First, this highlights the potential discrepancies between implementation details or hyperparameter settings that, due to stochasticity, can not be exactly replicated. Secondly, since our FAIM model retains and increases the importance of these features while still displaying fairness improvements, this could indicate that these features are intrinsically relevant to the task despite being potentially correlated to sensitive attributes. Further investigation into the conditions under which FAIM preserves or excludes these correlated attributes could be illuminating.

TABLE 5.2: Fairness metrics across intersectional groups (Sex × Race) for Baseline and FAIM models for IMV Task

Model	Metric	Min Intersection	Min Value	Max Intersection	Max Value	Gap	@Male_@White Value	@Male_@White Gap
Baseline	Equalized Odds (max of TPR/FPR)	Female_Black	0.6815	@Male_Asian	0.8933	0.2118	0.8564	0.1749
	Equal Opportunity (TPR)	Female_Black	0.6815	@Male_Asian	0.8933	0.2118	0.8564	0.1749
	BER Equality	@Male_Asian	0.2178	@Male_Other	0.2922	0.0744	0.2456	0.0278
FAIM	Equalized Odds (max of TPR/FPR)	Female_Black	0.6879	Female_Asian	0.8780	0.1902	0.8164	0.1285
	Equal Opportunity (TPR)	Female_Black	0.6879	Female_Asian	0.8780	0.1902	0.8164	0.1285
	BER Equality	Female_Other	0.2189	Female_Black	0.2832	0.0643	0.2409	0.0220

TABLE 5.3: Comparison of fairness and performance metrics across all bias mitigation methods

Method	Equal Opp.	Eq. Odds	Stat. Parity	Acc. Equality	BER Equality	Sensitivity	Specificity	AUC
Baseline	0.212	0.212	0.239	0.098	0.074	0.812	0.705	0.830
FAIM	0.190	0.190	0.146*	$0.075^{\dagger}$	0.064	0.788	0.726	0.828
Unawareness	0.195	0.195	0.152	0.083	0.063	0.807	0.711	0.829
Reweighing	0.170	0.179	0.191	0.097	0.081	0.808	0.707	0.828
Reductions	0.242	0.242	0.114	0.111	0.114	0.500	0.897	0.698
Equalized Odds	0.854	0.854	0.467	0.152	0.268	0.367	0.861	0.614
Calibrated Eq. Odds	0.780	0.780	0.456	0.203	0.259	0.417	0.879	0.648
Reject Option Classifier	0.188	0.188	0.190	0.083	0.070	0.785	0.729	0.757
Adversarial Learning	0.256	0.436	0.250	0.079	0.073	0.772	0.765	0.848

* Indicates FAIM significantly outperforms Baseline on Statistical Parity (p=0.024); † indicates FAIM significantly outperforms several fairness methods on Accuracy Equality (p=0.018)

Chapter 6

Conclusions and future work

6.1 Conclusion

In this study, we proved that the in-processing fairness method FAIM avoids performance degradation while improving fairness and maintaining interpretability in two distinct clinical tasks: predicting hospital admission and predicting invasive medical ventilation (IMV) allocation using the MIMIC-IV database. FAIM held up as a fairness method that preserves performance and interpretability while enhancing fairness metrics comprehensively in comparison to the fairness-unaware baseline logistic regression model. Even when outperformed by traditional fairness methods in one area, it surpassed them in others, illustrating its all-around effectiveness. It demonstrated stronger bias mitigation capabilities in a highly biased context, the original task, than it did in a more subtly biased context, the new task, but regardless improved fairness metrics in both scenarios. This discrepancy could also have been due to the large difference between dataset sizes used for either task. Stress-testing could still be deepened, especially by choosing a different, less transparent baseline model. Ultimately, FAIM held up its vigor with a different clinical task.

6.2 Limitations and Future Work

Dataset Expansion In this study, we only use information from MIMIC-IV, which stores data collected from just one part of the United States. The results obtained from training models on geographically limited data are not comprehensive for universal improved care. For the clinical task of IMV allocation, only patients admitted to the ICU were considered, which reduced the sample size greatly and didn't take into account other patients recorded as having IMV, for example those staying in the ED. For future exploration, it would be imperative to train FAIM models on larger, more diverse datasets.

Baseline Model Change The baseline model logistic regression was chosen for its transparency as it aligns with the transparency required in fields full of high-stakes decisions, in this case healthcare. Since FAIM is adaptable to other machine learning models [1], it would be interesting, and arguably even more relevant given the rapid development of deep learning models, to see if a more complex and less transparent model would hold up using FAIM. Due to the way that nearly-optimal models are generated in FAIM using ShapleyVIC, which is designed for generalized linear models (GLM), a distinct way of finding nearly optimal models would have to be explored. Hypothetically, this could be achieved using a sort of stochastic-training Rashomon set, in which nearly-optimal models are generated with varied dropout rates and diverse stopping criteria, retaining FAIM's threshold of nearly-optimal models remaining within 0.05 of the optimal model's AUC.

Sensitive Attribute Diversification Sensitive variables were kept to sex and race. The study could benefit from further expanding to consider other relevant sensitive features, such as marital status, insurance, and language spoken, which are all attributes included in the MIMIC-IV database. However, this would definitely require dataset expansion to address the challenge of small intersectional group sizes that currently limit FAIM applicability.

Appendix A

Supplementary Tables and Figures

TABLE A.1: MIMIC-IV Modules and Tables Used in Data Extraction

Module	Tables Used	Purpose
hosp	*patients, admissions, diagnoses_icd, labevents*	Demographics, admission info, comorbidities
icu	*icustays, chartevents, procedureevents*	ICU stay records, vital signs
derived	*ventilation, SOFA, CCI*	Ventilation status, severity scores, comorbidity index
ed	*edstays, triage, vitalsign, pyxis, medrecon*	Core ED data, vital signs, medications (for FAIM replication)

TABLE A.2: Dataset Split and Hospital Admission Distribution

Split	N	Percentage	Hospital Admission Count	Hospital Admission Rate
Overall	418,025	100.0%	197,799	47.32%
Training	292,617	70.0%	138,517	47.34%
Validation	41,802	10.0%	19,811	47.39%
Test	83,606	20.0%	39,471	47.21%

TABLE A.3: Dataset Split and Invasive Mechanical Ventilation (IMV) Distribution

Split	N	Percentage	IMV Count	IMV Rate
Overall	56,150	100.0%	23,644	33.53%
Training	39,305	70.0%	13,198	33.58%
Validation	5,615	10.0%	1,895	33.75%
Test	11,230	20.0%	3,732	33.23%

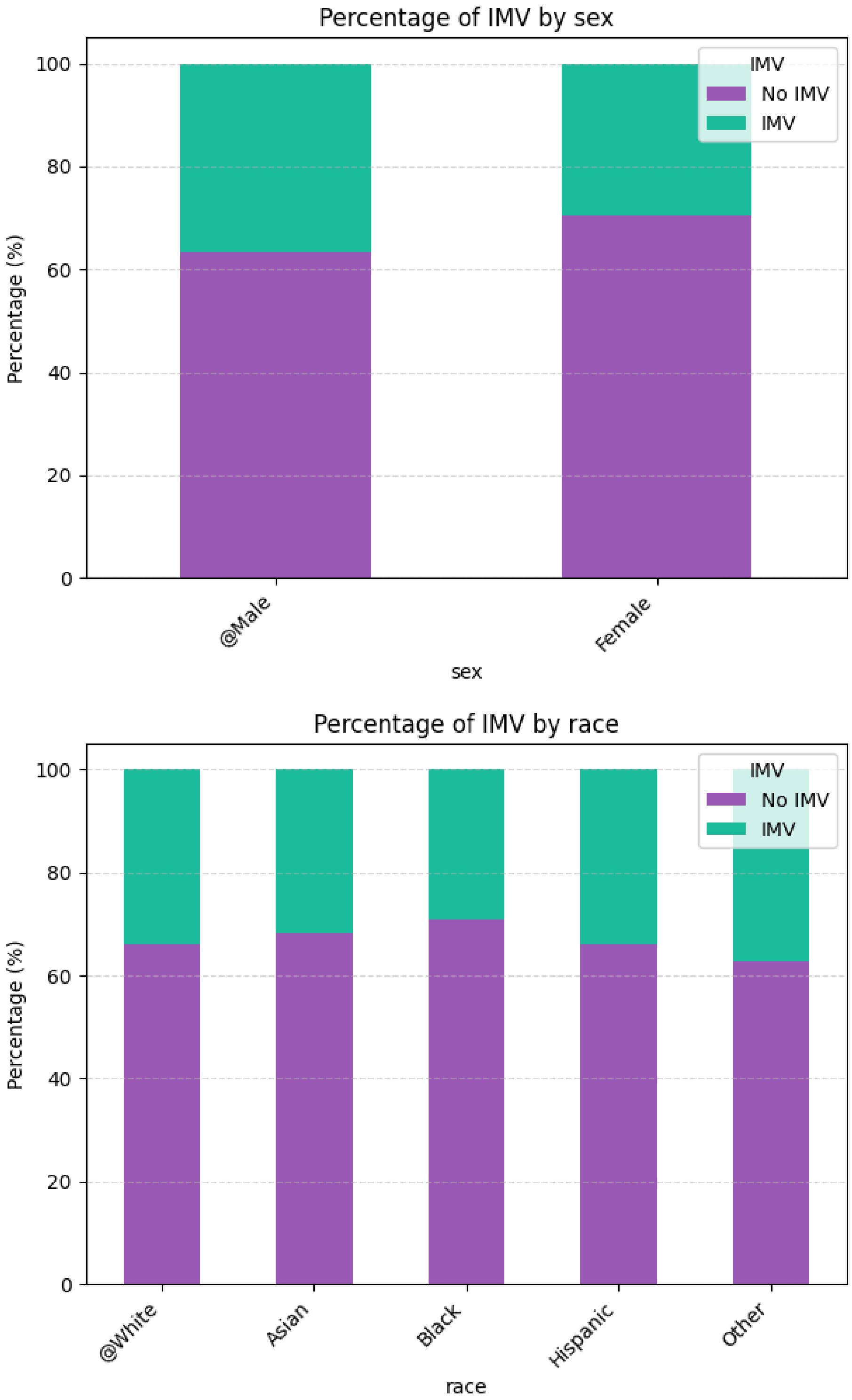


FIGURE A.1: Percentage of IMV by Sex and Race

TABLE A.4: Dataset Comorbidities

Feature	**Value**
Sample Size (N)	56,150
IMV Rate (%)	33.53
Age (years, mean ± std)	65.2 ± 16.9
Elective Admissions (%)	3.66
Comorbidities (%)	
Hypertension	33.91
Congestive Heart Failure	12.77
COPD	6.14
Asthma	3.97
Coronary Artery Disease	15.61
Chronic Kidney Disease	10.48
Diabetes	15.20
Connective Tissue Disease	0.98
Severity Scores (mean ± std)	
SOFA Score	4.57 ± 3.60
Charlson Comorbidity Index	4.81 ± 3.06
Vital Signs (First 24h Avg., mean ± std)	
Heart Rate (bpm)	84.1 ± 16.0
Temperature (°C)	36.82 ± 0.60
Respiratory Rate (bpm)	19.0 ± 3.9
Systolic BP (mmHg)	118.3 ± 17.8
Diastolic BP (mmHg)	65.9 ± 13.4

TABLE A.5: Mapping of MIMIC-IV Race Labels to Final Race Groups

Final Race Group	**Original MIMIC-IV Race Label**
White	WHITE WHITE - OTHER EUROPEAN WHITE - BRAZILIAN WHITE - RUSSIAN WHITE - EASTERN EUROPEAN
Black	BLACK/AFRICAN AMERICAN BLACK/AFRICAN BLACK/CAPE VERDEAN BLACK/CARIBBEAN ISLAND
Asian	ASIAN ASIAN - CHINESE ASIAN - ASIAN INDIAN ASIAN - SOUTHEAST ASIAN ASIAN - KOREAN
Hispanic	HISPANIC/LATINO - COLUMBIAN HISPANIC/LATINO - DOMINICAN HISPANIC/LATINO - PUERTO RICAN HISPANIC OR LATINO HISPANIC/LATINO - GUATEMALAN HISPANIC/LATINO - HONDURAN HISPANIC/LATINO - SALVADORAN HISPANIC/LATINO - CENTRAL AMERICAN HISPANIC/LATINO - CUBAN HISPANIC/LATINO - MEXICAN
Other	OTHER MULTIPLE RACE/ETHNICITY NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER AMERICAN INDIAN/ALASKA NATIVE SOUTH AMERICAN PORTUGUESE
Unknown	PATIENT DECLINED TO ANSWER UNKNOWN UNABLE TO OBTAIN

TABLE A.6: Separation-Based Fairness Metrics

Fairness Metric	**Formula**
Equalized Odds	max(Range(TPR), Range(FPR))
Equal Opportunity	Range(TPR)
Balanced Error Rate Equality	Range(0.5(FPR + FNR))
Statistical Parity	Range(SR)

Note: Range refers to the difference between maximum and minimum values across subgroups. Smaller values indicate better fairness and fewer biases. These metrics are limited to binary outcomes.

TABLE A.7: Component Metrics

Metric	**Formula**	**Description**
TPR (True Positive Rate)	TP / (TP + FN)	The proportion of actual positives that are correctly identified by the model within each subgroup. Also known as sensitivity or recall.
TNR (True Negative Rate)	TN / (TN + FP)	The proportion of actual negatives that are correctly identified by the model. Also known as specificity.
FPR (False Positive Rate)	FP / (FP + TN)	The proportion of actual negatives that are incorrectly classified as positive by the model within each subgroup.
FNR (False Negative Rate)	FN / (TP + FN)	The proportion of actual positives that are incorrectly classified as negative by the model within each subgroup.
SR (Selection Rate)	(TP + FP) / (TP + FP + TN + FN)	The proportion of all cases that are predicted as positive by the model within each subgroup. Also called positive prediction rate.

Note: Confusion Matrix Components: TP (True Positives): Correctly predicted positive cases, TN (True Negatives): Correctly predicted negative cases, FP (False Positives): Negative cases incorrectly predicted as positive, FN (False Negatives): Positive cases incorrectly predicted as negative.

TABLE A.8: Fairness methods

Stage	Method	Description
Pre-processing	Reweighing	Adjusts instance weights in the training data to reduce discrimination by rebalancing the influence of different demographic groups before model training.
	Unawareness	Removes sensitive attributes (e.g., race, sex) from the training data entirely, preventing the model from directly using them in predictions.
In-processing	FAIM	Generates nearly-optimal models using the Rashomon Effect and selects the fairness-optimal model based on Fairness Ranking Index (FRI), which aggregates equalized odds, equal opportunity, and BER equality.
	Reductions	Formulates fair classification as a constrained optimization problem, directly incorporating fairness constraints (e.g., equalized odds) during model training.
	Adversarial Learning	Uses adversarial neural networks where one network predicts the outcome and another tries to predict the sensitive attribute, forcing the predictor to remove demographic information.
Post-processing	Equalized Odds	Adjusts prediction thresholds after training to achieve equal TPR and FPR across demographic groups.
	Calibrated Equalized Odds	Similar to equalized odds but maintains calibration, ensuring predicted probabilities match actual outcome rates within groups.
	Reject Option Classifier	Creates a confidence band around the decision boundary where predictions are withheld or adjusted to improve fairness metrics.

TABLE A.9: Hospital Admission Task: Fairness Metrics Across Intersectional Demographic Subgroups for Baseline and FAIM models

Model	Metric	Min Intersection	Min Value	Max Intersection	Max Value	Gap
Baseline	Equalized Odds (max of TPR/FPR)	Female_Hispanic	0.4836	@Male_@White	0.8122	0.3286
	Equal Opportunity (TPR)	Female_Hispanic	0.4836	@Male_@White	0.8122	0.3286
	BER Equality	@Male_Others	0.2496	Female_Hispanic	0.3218	0.0721
FAIM	Equalized Odds (max of TPR/FPR)	Female_Asian	0.2163	@Male_@White	0.3604	0.1441
	Equal Opportunity (TPR)	Female_Hispanic	0.6236	@Male_@White	0.7612	0.1376
	BER Equality	@Male_Others	0.2536	Female_@White	0.3031	0.0495

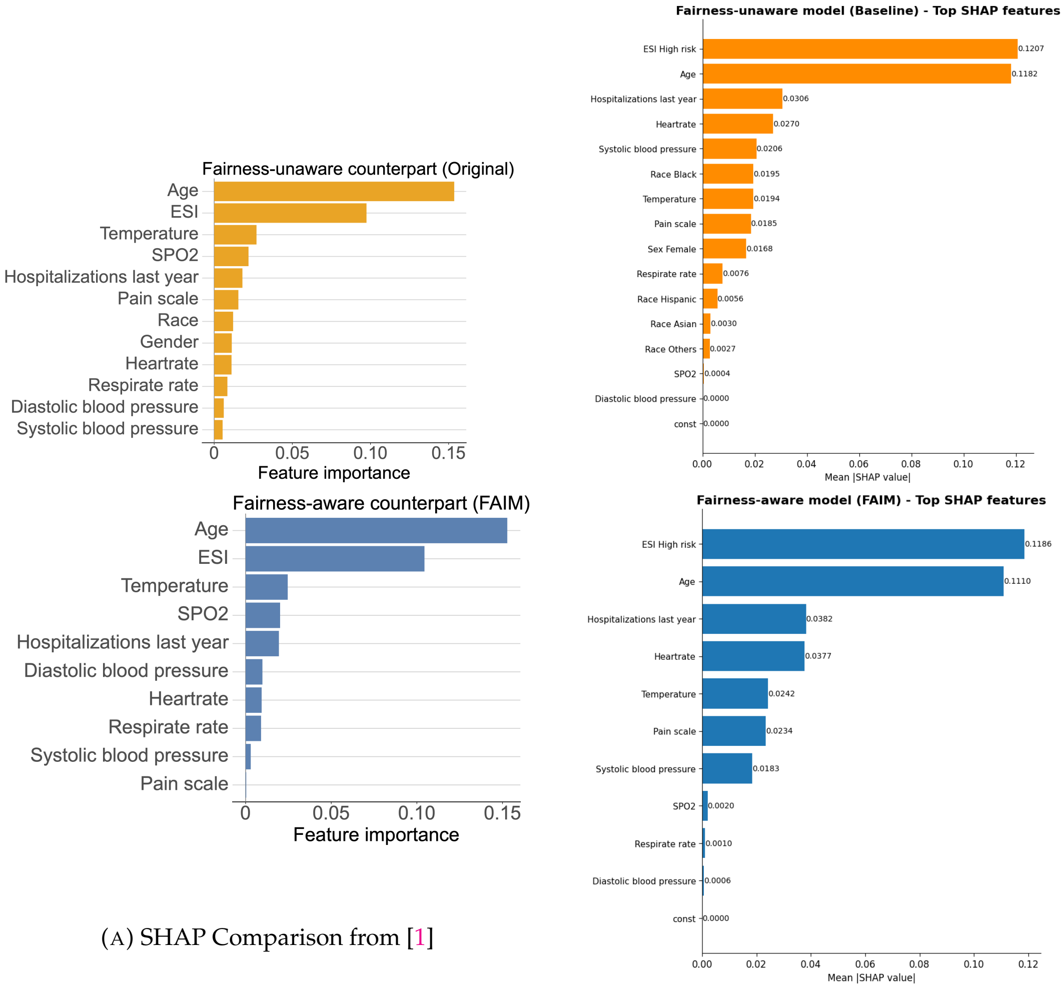
TABLE A.10: Hospital Admission Task: Comparison of Fairness Metrics Between Baseline and FAIM Models across Sex Subgroups

Method	Metric	Min Group	Min Value	Max Group	Max Value	Gap
Baseline	Equalized Odds (max of TPR/FPR)	Female	0.2541	@Male	0.3380	0.0840
	Equal Opportunity (TPR)	Female	0.6864	@Male	0.7599	0.0735
	BER Equality	Female	0.2838	@Male	0.2891	0.0052
FAIM	Equalized Odds (max of TPR/FPR)	Female	0.2894	@Male	0.3150	0.0256
	Equal Opportunity (TPR)	Female	0.7172	@Male	0.7424	0.0252
	BER Equality	Female	0.2861	@Male	0.2863	0.0002

TABLE A.11: Hospital Admission Task: Comparison of Fairness Metrics Between Baseline and FAIM Models across Race Subgroups

Method	Metric	Min Group	Min Value	Max Group	Max Value	Gap
Baseline	Equalized Odds (max of TPR/FPR)	Hispanic	0.5331	@White	0.7812	0.2481
	Equal Opportunity (TPR)	Hispanic	0.5331	@White	0.7812	0.2481
	BER Equality	Others	0.2537	Hispanic	0.3087	0.0550
FAIM	Equalized Odds (max of TPR/FPR)	Others	0.2223	@White	0.3519	0.1296
	Equal Opportunity (TPR)	Hispanic	0.6400	@White	0.7501	0.1101
	BER Equality	Others	0.2543	@White	0.3009	0.0466

FIGURE A.2: SHAP Comparison for Hospital Admission Task



(A) SHAP Comparison from [1]

(B) SHAP Comparison Replication

TABLE A.12: Distribution of Generated Models Across Exclusion Scenarios by Fairness Ranking Index (FRI)

Exclusion Cases	**1–10 ("most fair")**	**11–700**	**701–800**	**Highest Ranking**
Exclusion of sex and race	7	193	0	1
Exclusion of sex	2	181	17	2
Exclusion of race	1	183	16	6
No exclusion	0	133	67	83

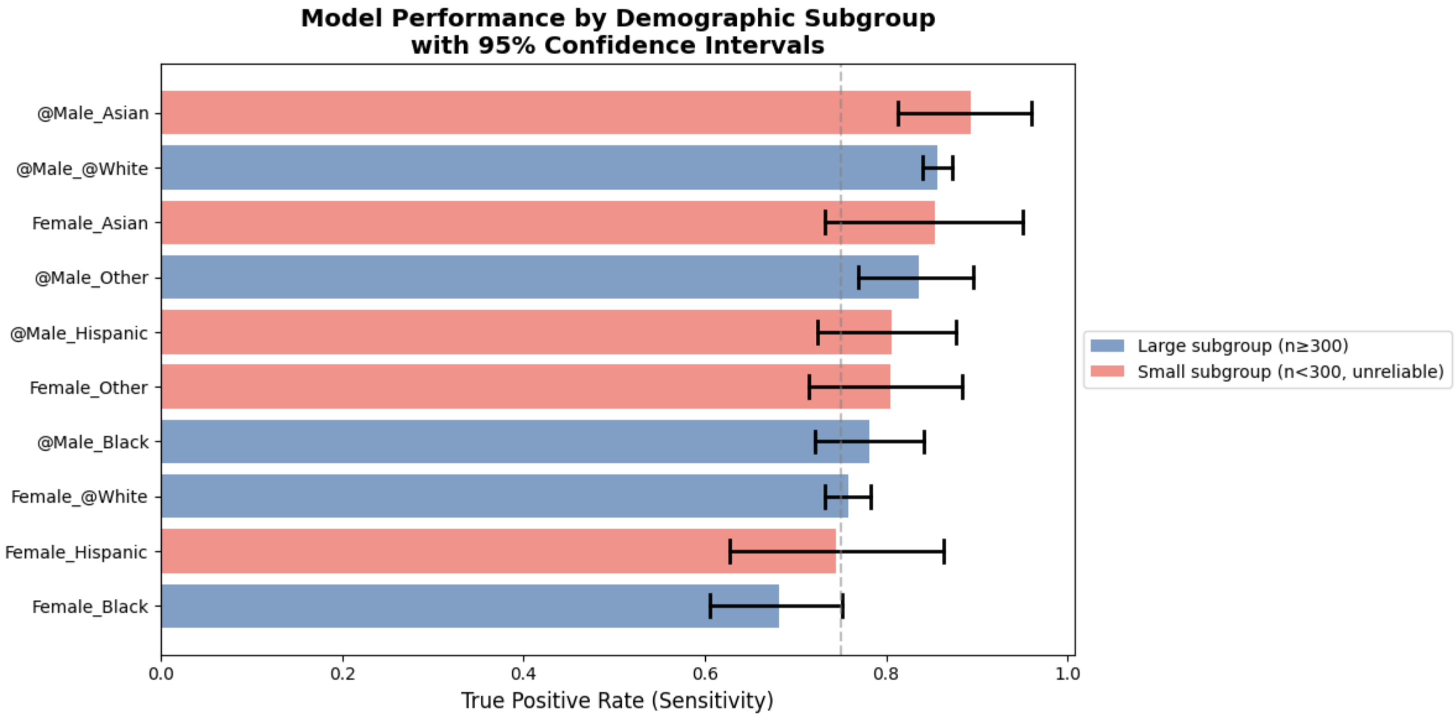


FIGURE A.3: Confidence Intervals across Demographic Subgroups

TABLE A.13: IMV Task: Comparison of Fairness Metrics Between Baseline and FAIM Models across Race Subgroups

Method	Metric	Min Group	Min Value	Max Group	Max Value	Gap	@White Value	@White Gap
Baseline	Equalized Odds (max of TPR/FPR)	Black	0.7353	Asian	0.8793	0.144	0.8186	0.0833
	Equal Opportunity (TPR)	Black	0.7353	Asian	0.8793	0.144	0.8186	0.0833
	BER Equality	Asian	0.2248	Other	0.2618	0.037	0.2403	0.0156
FAIM	Equalized Odds (max of TPR/FPR)	Black	0.7294	Asian	0.8448	0.1154	0.7946	0.0652
	Equal Opportunity (TPR)	Black	0.7294	Asian	0.8448	0.1154	0.7946	0.0652
	BER Equality	Asian	0.2301	Black	0.2612	0.0311	0.2411	0.0110

TABLE A.14: IMV Task: Comparison of Fairness Metrics Between Baseline and FAIM Models across Sex Subgroups

Method	Metric	Min Group	Min Value	Max Group	Max Value	Gap
Baseline	Equalized Odds (max of TPR/FPR)	Female	0.2451	@Male	0.3408	0.0957
	Equal Opportunity (TPR)	Female	0.7543	@Male	0.8483	0.0939
	BER Equality	Female	0.2454	@Male	0.2463	0.0009
FAIM	Equalized Odds (max of TPR/FPR)	Female	0.7550	@Male	0.8080	0.0530
	Equal Opportunity (TPR)	Female	0.7550	@Male	0.8080	0.0530
	BER Equality	@Male	0.2422	Female	0.2497	0.0074

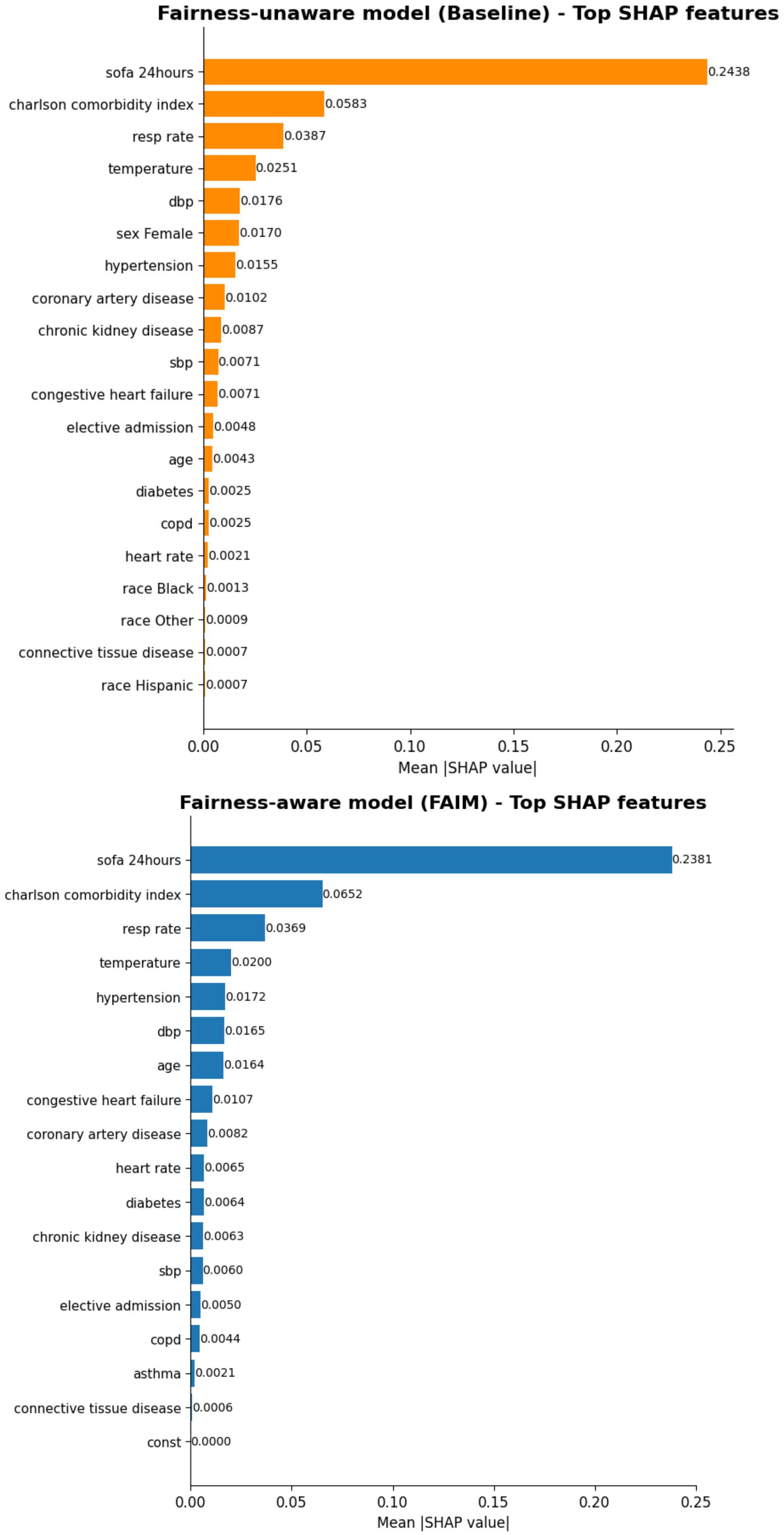
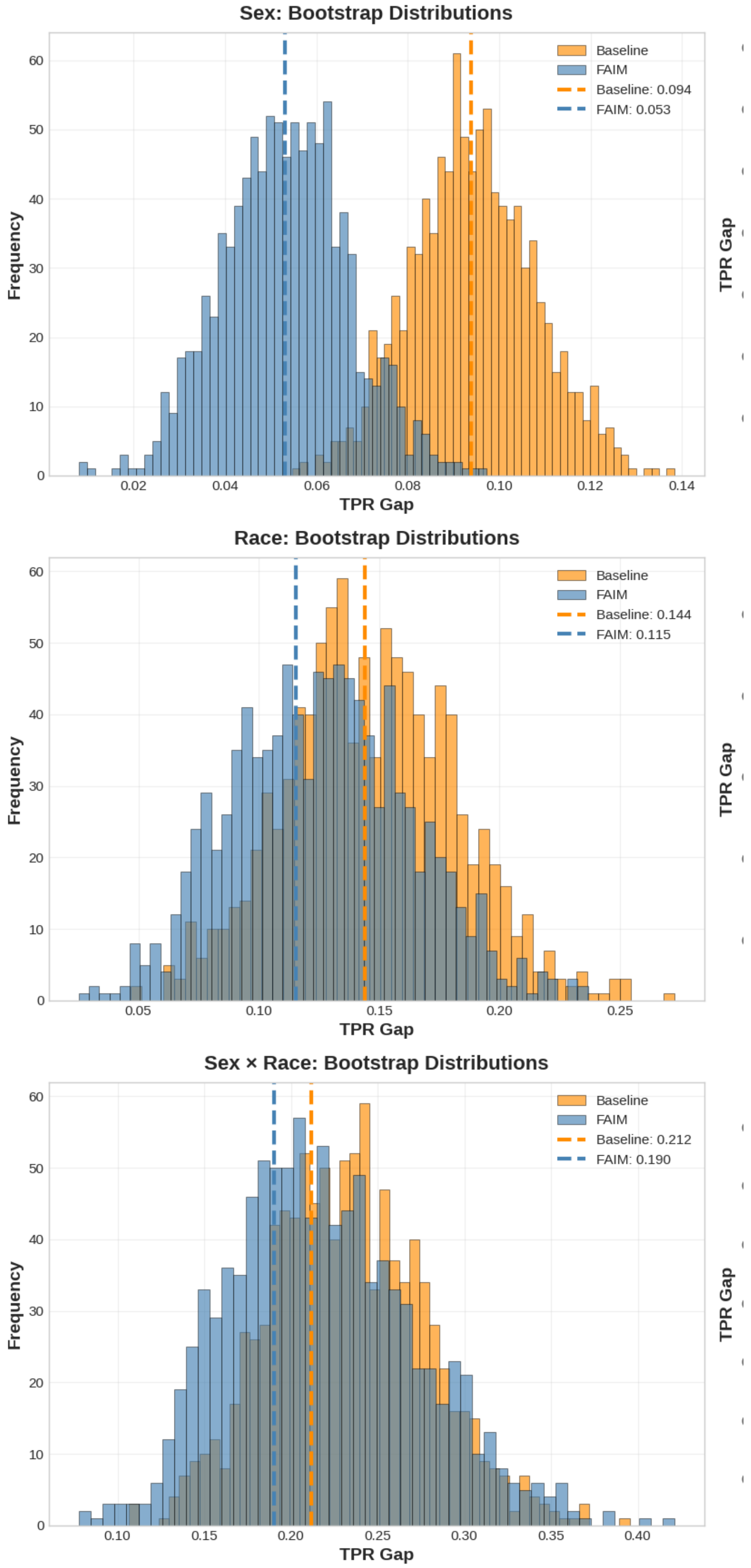


FIGURE A.4: SHAP Comparison for IMV Task



Sex: Bootstrap Distributions

Race: Bootstrap Distributions

Sex × Race: Bootstrap Distributions

FIGURE A.5: Bootstrap Distributions for IMV Task. Dashed vertical lines indicate mean TPR gaps. For sex, FAIM significantly reduces the gap (0.053 vs 0.094, p<0.001). For race and sex×race, gap reductions are not statistically significant, with substantial overlap between distributions.

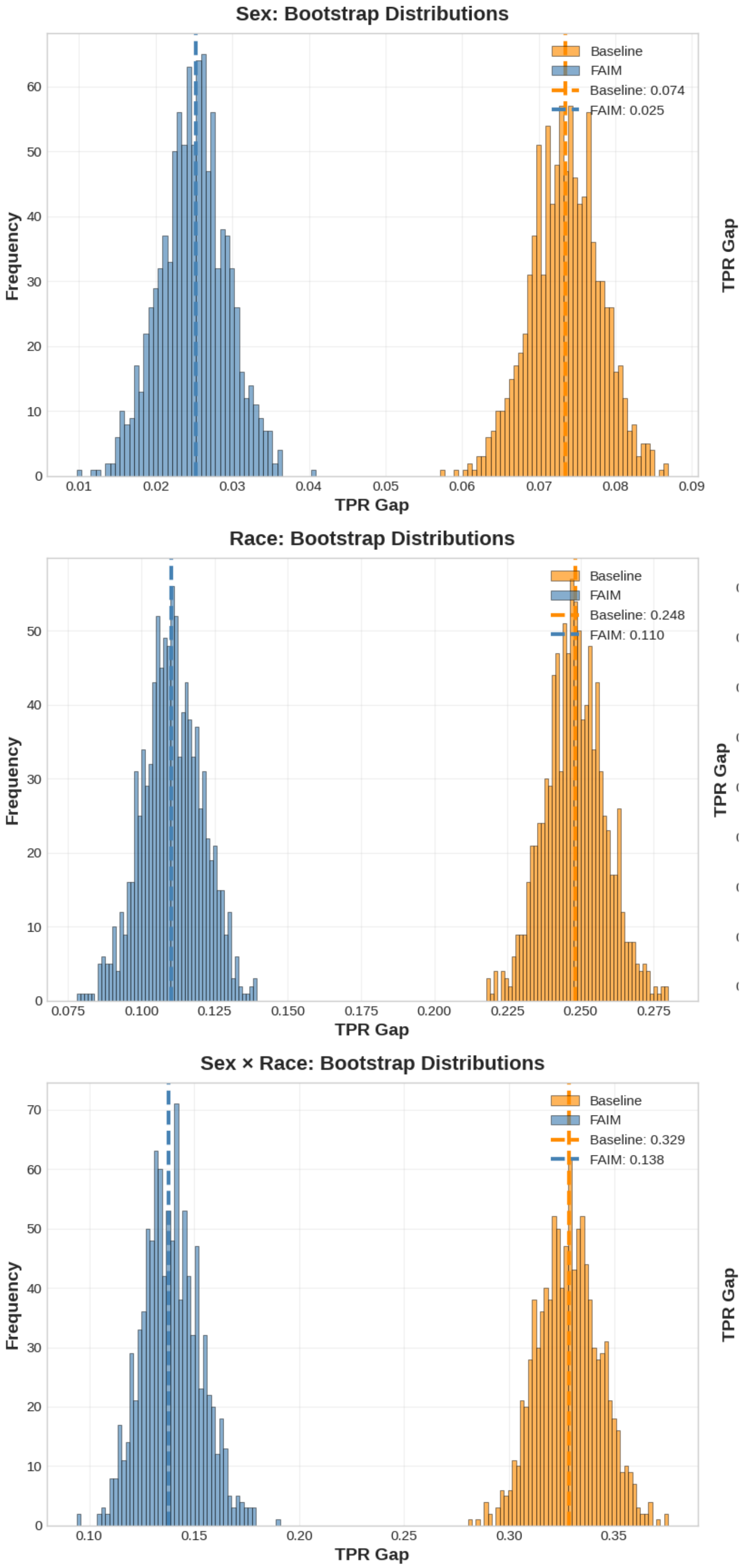
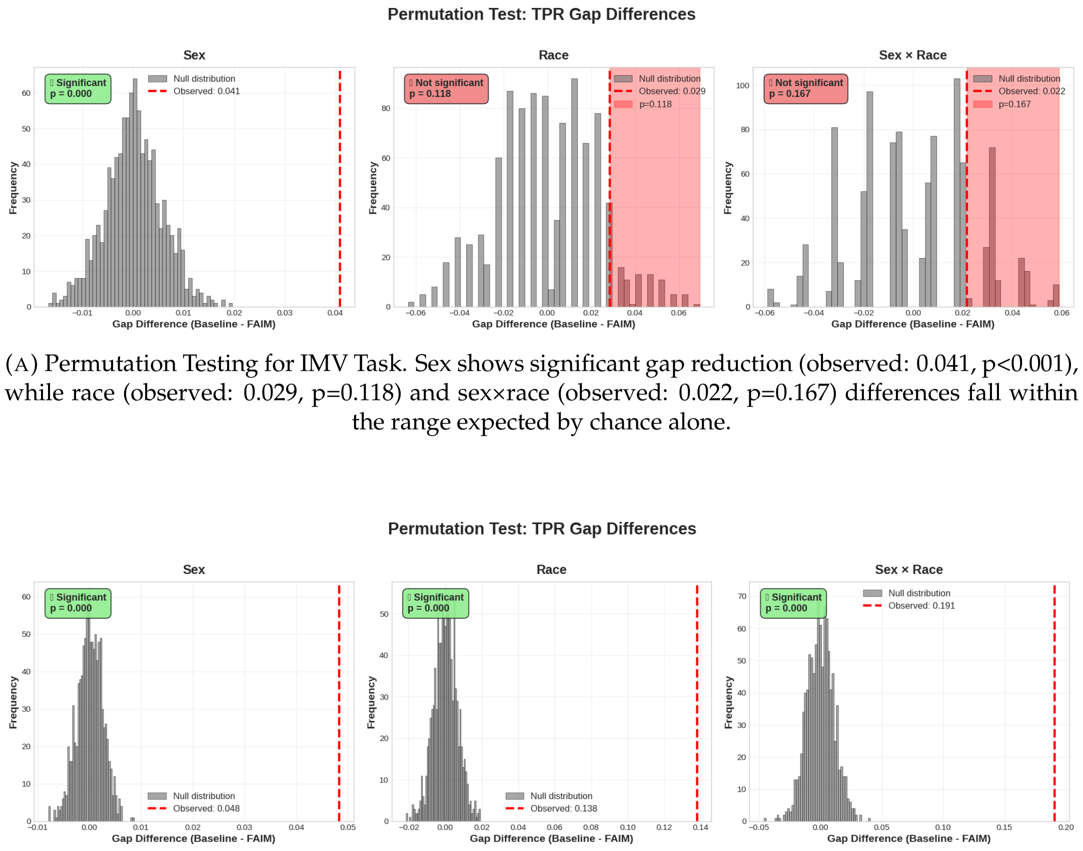


FIGURE A.6: Bootstrap Distributions for Hospital Admission Task. FAIM shows statistically significant reductions in TPR gaps compared to baseline across all demographic categories: sex (0.025 vs 0.074), race (0.110 vs 0.248), and sex×race intersections (0.138 vs 0.329).



(A) Permutation Testing for IMV Task. Sex shows significant gap reduction (observed: 0.041, p<0.001), while race (observed: 0.029, p=0.118) and sex×race (observed: 0.022, p=0.167) differences fall within the range expected by chance alone.

(B) Permutation Testing for Hospital Admission Task. All observed gap reductions fall outside their respective null distributions, indicating statistical significance: sex (observed: 0.048, p<0.001), race (observed: 0.138, p<0.001), and sex×race (observed: 0.191, p<0.001).

FIGURE A.7: Permutation Testing Results comparing baseline and FAIM model. Gray histograms represent null distributions of gap differences under random label permutation; red dashed lines show observed differences. Red boxes show non-significant results.

TABLE A.15: IMV Task: True Positive Rate Comparison across Intersectional Groups

Intersection	**N Total**	**N with IMV**	**TPR Baseline**	**TPR FAIM**	**TPR Change**	**Change %**
Female_Black	613	157	0.682	0.688	+0.006	+0.9%
Female_Hispanic	203	51	0.745	0.726	-0.020	-2.6%
Female_@White	3806	1119	0.758	0.760	+0.002	+0.2%
@Male_Black	565	183	0.781	0.765	-0.016	-2.1%
Female_Other	229	77	0.805	0.779	-0.026	-3.2%
@Male_Hispanic	267	98	0.806	0.765	-0.041	-5.1%
@Male_Other	341	134	0.836	0.776	-0.060	-7.1%
Female_Asian	184	41	0.854	0.878	+0.024	+2.9%
@Male_@White	4795	1797	0.856	0.816	-0.040	-4.7%
@Male_Asian	227	75	0.893	0.827	-0.067	-7.5%

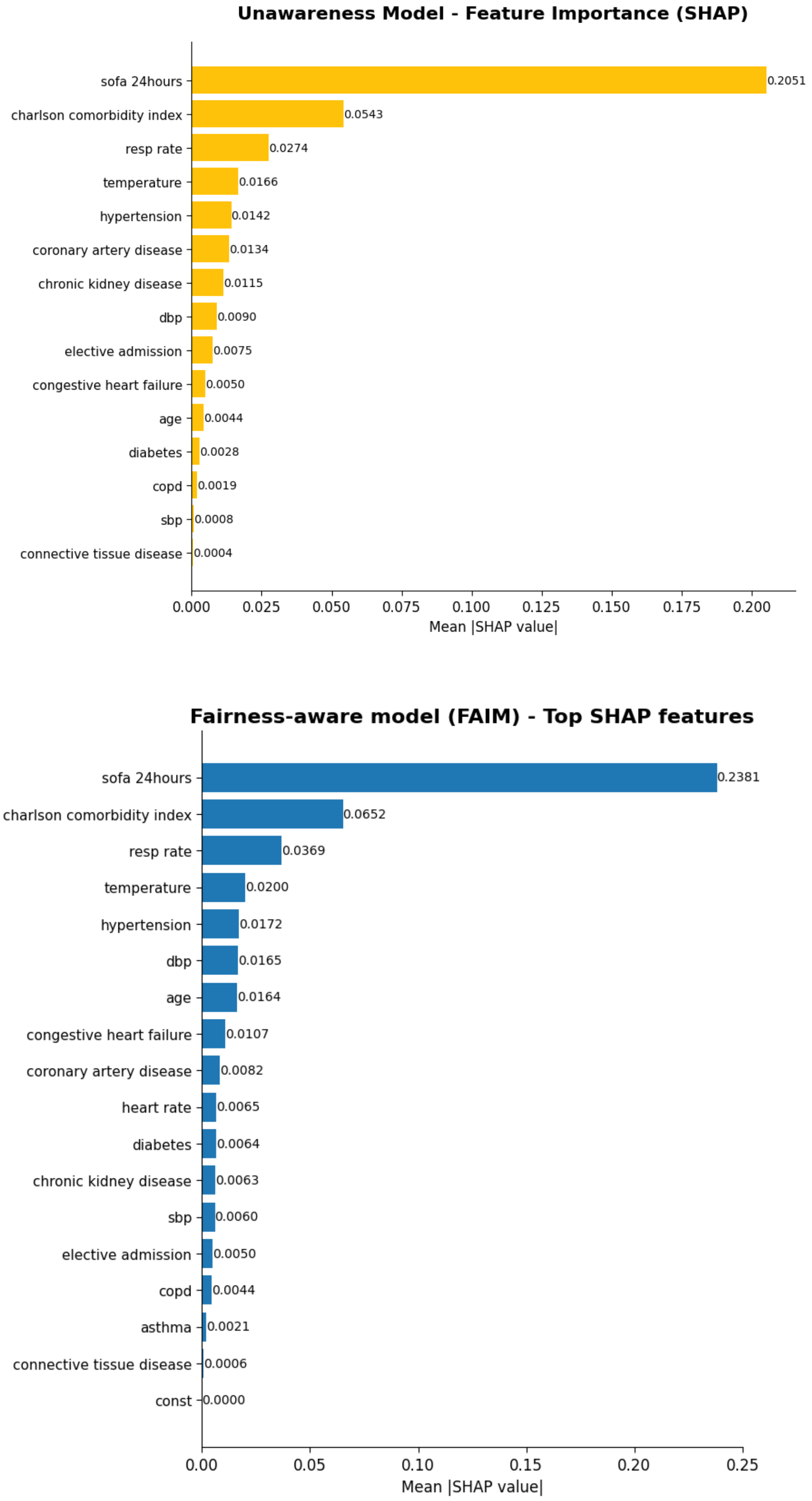
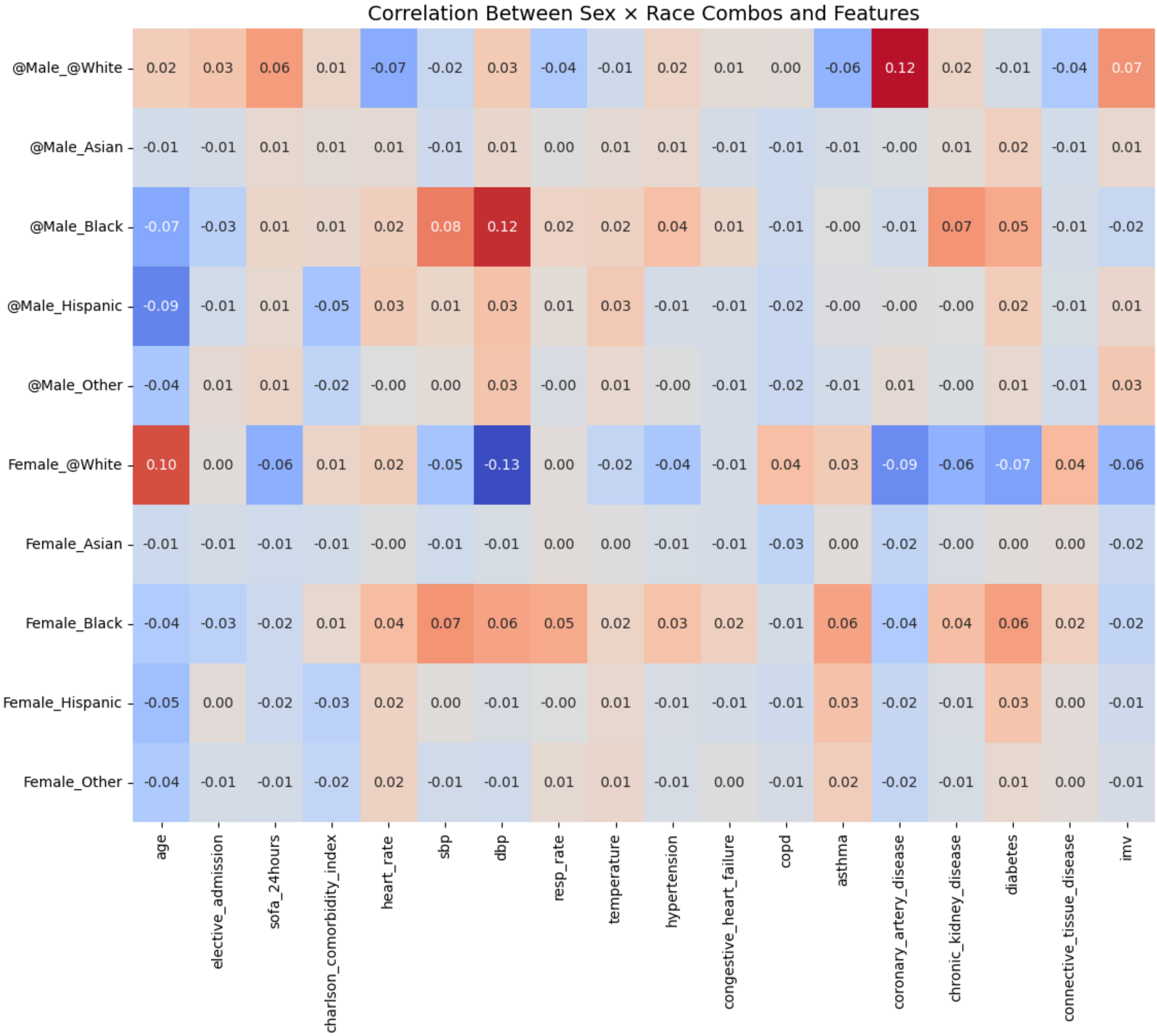


FIGURE A.8: SHAP Comparison Between Unawareness and FAIM for IMV Task



Correlation Between Sex × Race Combos and Features

	age	elective_admission	sofa_24hours	charlson_comorbidity_index	heart_rate	sbp	dbp	resp_rate	temperature	hypertension	congestive_heart_failure	copd	asthma	coronary_artery_disease	chronic_kidney_disease	diabetes	connective_tissue_disease	imv
@Male_@White	0.02	0.03	0.06	0.01	-0.07	-0.02	0.03	-0.04	-0.01	0.02	0.01	0.00	-0.06	0.12	0.02	-0.01	-0.04	0.07
@Male_Asian	-0.01	-0.01	0.01	0.01	0.01	-0.01	0.01	0.00	0.01	0.01	-0.01	-0.01	-0.01	-0.00	0.01	0.02	-0.01	0.01
@Male_Black	-0.07	-0.03	0.01	0.01	0.02	0.08	0.12	0.02	0.02	0.04	0.01	-0.01	-0.00	-0.01	0.07	0.05	-0.01	-0.02
@Male_Hispanic	-0.09	-0.01	0.01	-0.05	0.03	0.01	0.03	0.01	0.03	-0.01	-0.01	-0.02	-0.00	-0.00	-0.00	0.02	-0.01	0.01
@Male_Other	-0.04	0.01	0.01	-0.02	-0.00	0.00	0.03	-0.00	0.01	-0.00	-0.01	-0.02	-0.01	0.01	-0.00	0.01	-0.01	0.03
Female_@White	0.10	0.00	-0.06	0.01	0.02	-0.05	-0.13	0.00	-0.02	-0.04	-0.01	0.04	0.03	-0.09	-0.06	-0.07	0.04	-0.06
Female_Asian	-0.01	-0.01	-0.01	-0.01	-0.00	-0.01	-0.01	0.00	0.00	-0.01	-0.01	-0.03	0.00	-0.02	-0.00	0.00	0.00	-0.02
Female_Black	-0.04	-0.03	-0.02	0.01	0.04	0.07	0.06	0.05	0.02	0.03	0.02	-0.01	0.06	-0.04	0.04	0.06	0.02	-0.02
Female_Hispanic	-0.05	0.00	-0.02	-0.03	0.02	0.00	-0.01	-0.00	0.01	-0.01	-0.01	-0.01	0.03	-0.02	-0.01	0.03	0.00	-0.01
Female_Other	-0.04	-0.01	-0.01	-0.02	0.02	-0.01	-0.01	0.01	0.01	-0.01	0.00	-0.01	0.02	-0.02	-0.01	0.01	0.00	-0.01

FIGURE A.9: Correlation Matrix

Appendix B

Code Availability

The code developed throughout this study is publicly available in the following GitHub repository: https://github.com/irisvukovic/DS-Master-Thesis-UB

Bibliography

[1] M. Liu *et al.*, “FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare,” *Patterns*, vol. 5, Oct. 2024.

[2] F. M. Abdelmalek *et al.*, “Association between patient race and ethnicity and use of invasive ventilation in the United States,” *Ann. Am. Thorac. Soc.*, vol. 21, pp. 287–295, Feb. 2024.

[3] A. E. W. Johnson *et al.*, “MIMIC-IV, a freely accessible electronic health record dataset,” *Sci. Data*, vol. 10, Jan. 2023.

[4] G. B. Moody and R. G. Mark, “MIMIC database (version 1.0.0).” PhysioNet, Mar. 2000. Available: `https://physionet.org/content/mimic/1.0/`.

[5] A. Johnson *et al.*, “MIMIC-IV (version 3.1).” PhysioNet, Oct. 2024.

[6] PhysioNet, “PhysioNet: Research resource for complex physiologic signals.” Online. Available: `https://physionet.org/`.

[7] CITI Program, “CITI Program: Research ethics & compliance training.” Online. Available: `https://about.citiprogram.org/`.

[8] MIMIC, “MIMIC: Medical information mart for intensive care.” Online. Available: `https://mimic.mit.edu/`.

[9] MIT-LCP, “mimic-code: MIMIC-IV.” GitHub repository, 2025. Available: `https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iv`.

[10] V. Parcha *et al.*, “Trends and geographic variation in acute respiratory failure and ARDS mortality in the United States,” *Crit. Care*, vol. 159, pp. 1460–1472, Apr. 2021.

[11] C. Bime *et al.*, “Racial differences in mortality from severe acute respiratory failure in the United States, 2008–2012,” *Ann. Am. Thorac. Soc.*, vol. 13, pp. 867–875, June 2016.

[12] NliuLab, “FAIM: Fairness-aware interpretable modeling.” GitHub repository, 2025. Available: `https://github.com/nliulab/FAIM`.

[13] NliuLab, “mimic4ed-benchmark.” GitHub repository, 2022. Available: `https://github.com/nliulab/mimic4ed-benchmark`.

[14] AAMC Center for Health Justice, “Advancing health equity: A guide to language, narrative and concepts.” Online. Available: `https://www.aamchealthjustice.org/key-topics/trustworthiness/narrative-guide`.

[15] M. Rajunov *et al.*, eds., *Nonbinary: Memoirs of Gender and Identity*. New York, NY: Columbia University Press, Apr. 2019.

[16] S. Dev *et al.*, “Harms of gender exclusivity and challenges in non-binary representation in language technologies,” in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, (Punta Cana, Dominican Republic), pp. 1968–1994, Nov. 2021.

[17] L. Oneto *et al.*, “Fairness in machine learning,” in *Recent Trends in Learning From Data*, vol. 896 of *Stud. Comput. Intell.*, Cham: Springer, 2020.

[18] J. Gao *et al.*, “What is fair? Defining fairness in machine learning for health,” *Stat. Med.*, vol. 44, p. e70234, Sept. 2025.

[19] J. Yang *et al.*, “An adversarial training framework for mitigating algorithmic biases in clinical machine learning,” *npj Digit. Med.*, vol. 6, Mar. 2023.

[20] A. Castelnovo *et al.*, “A clarification of the nuances in the fairness metrics landscape,” *Sci. Rep.*, vol. 12, Mar. 2022.

[21] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” *Knowl. Inf. Syst.*, vol. 33, pp. 1–33, Oct. 2012.

[22] M. Hardt *et al.*, “Equality of opportunity in supervised learning.” arXiv:1610.02413 [cs.LG], Oct. 2016.

[23] A. Agarwal *et al.*, “A reductions approach to fair classification.” arXiv:1803.02453 [cs.LG], 2018.

[24] C. Rudin *et al.*, “Amazing things come from having many good models,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[25] Y. Ning *et al.*, “Shapley variable importance cloud for interpretable machine learning,” *Patterns*, vol. 3, Apr. 2022.

[26] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, pp. 32–35, Jan. 1950.

[27] S. Lundberg *et al.*, “A unified approach to interpreting model predictions.” arXiv:1705.07874 [cs.AI], 2017.