FairMed - XGB：一种经贝叶斯优化的面向人口统计学公平性的多指标可解释框架重症医疗数据中的人口统计学公平性

Mitul Goswami[1], 、Romit Chatterjee[1] 和 Arif Ahmed Sekh[2]

[1] 卡林加工业技术学院计算机工程学院，751024，印度布巴内斯瓦尔[2] 挪威北极大学生物人工智能实验室，挪威特罗姆瑟

摘要。 部署于重症监护场景中的机器学习模型存在人口统计偏差，尤其是性别差异，这会损害临床信任并妨碍公平治疗。本文提出 FairMed-XGB，这是一种新颖的框架，可系统地检测并缓解基于性别的预测偏差，同时保持模型性能和透明度。该框架将结合统计均等差异、泰尔指数和瓦瑟斯坦距离的公平性感知损失函数，集成到 XGBoost 分类器中，并通过贝叶斯搜索进行联合优化。在基于 MIMIC-IV-ED 和 eICU 数据库构建的七个临床特征各异的队列上进行缓解后评估，结果显示偏差显著减少：MIMIC-IV-ED 上的统计均等差异下降 40–51%，eICU 上下降 10–19%；泰尔指数降低四至五个数量级，趋近于零；瓦瑟斯坦距离减少 20–72%。这些提升几乎未造成预测准确性下降（AUC-ROC 降幅 <0.02）。基于 SHAP 的可解释性分析表明，该框架降低了对性别代理特征的依赖，为临床医生提供了关于偏差在何处以及如何得到纠正的可操作见解。FairMed-XGB 为实现公平的临床决策提供了一种稳健、可解释且符合伦理原则的解决方案，为人工智能在高风险医疗环境中的可信部署铺平道路。

关键词：公平机器学习、性别偏差缓解、贝叶斯优化、XGBoost、SHAP、重症监护、MIMIC-IV、eICU、算法公平性、可解释人工智能。

1 引言

机器学习（ML）已成为医疗保健领域的一股变革性力量，尤其是在重症监护病房（ICU）和急诊科等高风险重症监护场景中。先进的预测模型日益广泛地用于预测患者结局、优化资源分配并指导治疗方案，借助海量数据集提升临床决策水平 [1][2]。例如，基于电子健康记录（EHR）训练的算法能够以惊人的准确率预测脓毒症发病、死亡风险或再入院可能性，从而实现积极主动的干预——

，挽救生命并降低成本 [3]。这些模型有望大幅简化工作流程、减少人为错误并提升护理服务的公平性。然而，对数据驱动系统日益增长的依赖也凸显出一项紧迫的伦理挑战：训练数据或算法设计中固有的偏见，可能会延续或加剧弱势群体所面临的差异。研究表明，人口统计学方面的失衡，例如心血管数据集中女性患者代表性不足，或疼痛评估算法中的种族差异，往往会导致预测偏差、误诊或资源分配不公 [4][5]。在重症监护中，及时且不带偏见的决策至关重要；此类缺陷可能加深现有的医疗卫生不平等，并对边缘化群体造成不成比例的影响。因此，确保机器学习驱动的临床工具具备公平性，不仅是技术问题，更是维护公平、增进现代医疗保健体系信任的道德责任。

机器学习模型在医疗保健领域的有效性取决于训练数据的代表性和质量，然而，临床数据集中普遍存在的人口统计偏差往往会损害其公平性。偏差源于数据收集中的系统性不平等，例如少数群体的代表性不足、不同人口统计群体之间不平衡的结局，或社会经济因素记录不完整 [6, 7]。例如，使用性别比例失衡的重症监护病房数据集训练的模型，可能会系统性地错误估计女性患者的死亡风险；而利用种族构成单一数据的算法，则可能无法准确地对来自代表性不足背景的患者进行分诊 [8]。这些偏差会在预测流程中不断传播，进一步扩大关键结局方面的差异。一项于2022年开展的eICU数据研究显示，与白人患者相比，预测脓毒症的模型对黑人患者的召回率显著降低，从而加剧了救命干预措施的延误 [9]。同样，研究表明，分诊算法中嵌入的社会经济偏差会不成比例地将资源从低收入人群转移出去，进一步加重现有的医疗可及性障碍 [10]。此类失误不仅会削弱人们对临床人工智能的信任，还可能使不公平循环长期延续，使边缘化群体在急性护理环境中面临叠加的不利处境。应对这些挑战需要建立能够明确识别并缓解数据驱动偏差、同时保留模型实用性的框架——本研究旨在弥合这一空白。

现有的偏见缓解技术，例如重加权和对抗式去偏，往往优先关注孤立的公平性指标，如统计均等，却忽视了医疗保健中人口统计学公平性的多面性 [11, 12]。例如，重加权方法可能会平衡数据集中各组的代表性，却无法解决模型性能指标方面的差异，例如假阴性率，而这在危及生命的病症中至关重要。同样，对抗式方法若抑制敏感属性之间的相关性，可能会降低预测准确率，或忽视由人口因素相互重叠而产生的交叉偏差（例如种族和性别）[13]。此外，当前框架很少协调相互竞争的公平性目标，例如均等化赔率（错误率相等）和分布公平性（各组预测分布趋于一致），导致临床环境中的权衡并非最优 [14]。雪上加霜的是，公平性意识模型缺乏可解释性——

许多方法如同“黑箱”，无法让临床医生充分了解偏见缓解如何影响决策逻辑或特征重要性 [15]。这种不透明性妨碍了信任与应用，因为医疗保健提供者需要透明的工具，以根据伦理和临床标准验证模型行为。

本研究提出 FairMed -XGB，这是一种旨在通过三项协同创新提升重症医疗预测模型人口统计学公平性的新型框架。首先，该框架采用贝叶斯优化的多指标公平性，动态平衡均等化赔率、泰尔指数和瓦瑟斯坦距离，使准确性与亚组公平性相协调。不同于以往静态设定公平性约束的方法，该方法会自适应地调整超参数，以优化在不同重症监护病房数据集（如 eICU 和 MIMIC-IV ED）上的权衡。其次，该框架优先考虑人口统计学公平性，通过严格评估各亚组的预测分布和错误率，明确缓解临床结局中的性别差异——这是现有重症监护模型中的一个关键缺口。第三，FairMed – XGB 集成 SHAP，针对特征贡献提供细致且可供临床医生采取行动的见解，弥合公平性感知机器学习中的可解释性鸿沟。通过将 SHAP 值与偏见关联起来，该框架确保模型决策和偏见缓解策略均具备透明度。

2 相关研究

2.1 用于机器学习与深度学习的医疗表格数据

医疗表格数据集（如 MIMIC-IV ED [16] 和 eICU [17], ）对于训练重症监护领域的机器学习模型至关重要。这些数据集包含结构化电子健康记录，包括人口统计信息、生命体征和实验室结果，可用于脓毒症检测等预测任务。然而，固有偏差（例如少数群体代表性不足或结局不平衡）会损害模型的公平性。例如，MIMIC-IV ED 的心血管记录存在性别差异，而 eICU 数据反映出脓毒症预测召回率方面的种族不平衡。近期研究指出，这类数据集中普遍存在社会经济和种族偏差变量，因此有必要进行预处理或采取算法缓解措施。表 1 总结了常见数据集及相关偏差变量。

表 1. 医疗表格数据集中的偏差变量

数据集	描述	偏差变量
MIMIC-IV [16]	贝斯以色列医院的重症监护病房数据	性别、种族、年龄
eICU [17]	多中心 ICU 数据	种族、社会经济地位
NHANES [18]	国家健康调查数据	族裔、收入、教育
英国生物样本库 [19]	英国大型生物医学数据库	族裔、社会经济地位
All of Us [20]	美国国立卫生研究院资助的多样化队列数据	地理多样性，种族
弗雷明汉心脏研究 [21]	纵向心血管研究	历史队列
BRFSS [22]	美国疾病控制与预防中心 行为风险因素调查	自我报告偏倚，收入

2.2 医疗数据中的偏差检测方法

检测医疗数据集中的偏差，需要采用系统化方法来识别和量化不平等。差异影响分析 [23] 比较特权群体与边缘化群体获得有利结局的比例（例如按种族划分的重症监护病房收治率），以识别系统性差异。因果图分析 [24] 利用有向无环图（DAG）找出混杂变量（例如社会经济地位），并分离偏差产生的路径。对抗式去偏 [25] 通过对抗网络训练模型，以在学习过程中尽量减少偏差，但有时可能会降低预测准确率。公平性审计 [26] 使用假阳性率（FPR）和假阴性率（FNR）等指标评估子群体表现，从而揭示差异（例如黑人患者的脓毒症预测差异）。反事实测试 [27] 改变敏感属性（例如性别），以检查预测不变性并揭示因果偏差。重加权技术 [28] 调整数据权重（例如对女性心血管病例进行过采样），以解决表征失衡。最后，子群体表现分析 [29] 按人口统计学划分（例如年龄、族裔）评估模型表现，如 MIMIC-IV 研究 [30]所示。

2.3 医疗数据中的偏差缓解方法

医疗数据中的偏差缓解采用各种技术，以减少不同人口统计学群体在模型结果上的差异。重加权 [31] 通过调整样本权重，使代表性不足的队列达到平衡，例如在心血管数据集中对女性患者进行过采样。对抗式去偏 [32] 通过对抗网络训练模型，以尽量降低预测值与敏感属性（例如脓毒症风险模型中的种族）之间的相关性。公平性约束 [33] 将公平性指标（例如均等化赔率）纳入损失函数，在训练过程中惩罚歧视性预测。数据增强 [34], 包括合成少数类过采样（SMOTE），用于解决罕见病数据集中的类别不平衡问题。因果缓解方法 [35] 利用有向无环图对社会经济地位等混杂因素进行调整，以确保结果中的因果关系不受偏见影响。后处理技术 [36] 重新校准决策阈值（例如重症监护病房出院预测），以使不同亚组之间的错误率保持一致。公平元学习 [37] 通过在数据集之间迁移公平性感知知识，使模型适应异质人群。尽管这些方法能够减少偏见，但在保持临床准确性以及解决交叉性差异（例如种族与性别交互作用）方面仍面临挑战。

2.4 可解释的偏差检测与缓解方法

可解释方法提升了医疗人工智能中识别和解决偏差的透明度。SHAP（SHapley 加性解释）[38] 量化特征对预测值的贡献，从而揭示种族或保险类型等变量中的偏差。LIME（局部可解释模型无关解释）[39] 生成局部解释，以凸显有偏的决策逻辑（例如 eICU 数据中的分诊差异）。FairML [40] 通过输入扰动审查模型，将偏差追溯到具体特征（例如，邮政编码

是种族的代理变量）。可解释公平性约束[41],（如均等化赔率惩罚）将公平性指标纳入模型训练，同时提供决策边界透明度。反事实解释[42] 通过测试敏感属性发生假设性变化时预测值是否保持不变（例如，在死亡风险模型中改变性别），来识别因果偏差路径。基于规则的模型[43],（如公平性感知决策树）会执行可解释公平性规则（例如，公平的 ICU 出院标准）。

3 方法论

本节详细介绍 FairMed - XGB 框架。这是一种系统性方法，旨在检测并缓解临床机器学习模型中的基于性别的预测偏差，同时保持预测性能和可解释性。该方法论包含四个核心阶段：(1) 数据预处理，(2) 缓解前的偏见检测与分析，(3) 公平性感知损失函数的制定与整合，以及 (4) 缓解后训练与评估。该框架的数学公式见算法1。

算法1：FairMed – XGB：贝叶斯优化的多指标公平性学习 输入：临床数据集 $D = \{(x_i, y_i, a_i)\}_{i=1}^{n}$，其中$x_i \in \mathbb{R}^d$：特征向量 $y_i \in \{0,1\}$：二元临床标签 $a_i \in \{0,1\}$：敏感属性（性别） 输出：公平性感知模型 f^* 数据预处理 对分类特征进行编码：$x_{i,j} \leftarrow \varphi(x_{i,j})$ 标准化连续特征：$x_{i,j} \leftarrow (x_{i,j} - \mu_j)/\sigma_j$ 将D 划分为D_{Train} 和D_{Test} （按a_i分层抽样） 基线训练 初始化模型：$f(x) = 0$ ()至T do计算预测值：$\hat{y}_i = \sigma(f(x_i)) = \frac{1}{(1 + e^{-f(x_i)})}$ 计算梯度：$For\ t = 1$ $g_i = \frac{\delta L_{Pred}}{\delta (f(x_i))}$

将回归树 h$_t(x)$ 拟合到 $\{g_i\}$ 更新模型：$f(x) \leftarrow f(x)+ n_t$ h(x) $End\ For$偏差量化 定义组：$G_0 = \{i \mid a_i = 0\}$ $G_1 = \{$ $i \mid a_i = 1\}$ 计算统计均等差异：$SPD = P(\hat{Y} = 1 \mid A = 1) - P$ $(\hat{Y} = 1 \mid A = 0)$ 计算泰尔指数：$Theil = \sum_{i=1}^{m} \frac{\hat{p}_i}{p} \ln($ $\frac{\hat{p}_i}{p}$ $w\ ere\ \ddot{p} = \frac{1}{m}\sum_{i=1}^{m} \hat{p}_i$)，h 计算瓦瑟斯坦距离：

$W = \int_0^1 |F_1(t) - F_0(t)|dt$ 公平性感知 目标 定义公平性损失：L$_{fair} = w_1 \cdot SPD + w_2 \cdot Th$ $eil + w_3 \cdot W$ 定义总损失：L$_{\text{total}}$ = L$_{log} + \lambda \cdot ($ L$_{fair}$) 贝叶斯优化 初始化搜索空间：$\theta = (\lambda, w_1, w_2, w_3)$ $For\ k = 1\ to\ K\ do$训练模型以最小化 L

$_{\text{total}}$ 评估验证目标：

$J(\theta) = \alpha * AUC - (1-\alpha) * {}_{fair}$ () ( ) L 使用高 θ 斯过程代理模型进行更新 $End\ For$ 选择最优的θ^* 最终训

练 使用θ^* 在D_{train} 上重新训练模型 获得最终模型f^* 可解 释性 For $j\ do$ 每个特征 计算值 $SHAP$ φ_j 计算组间差异：

$\Delta\varphi_j = \mathrm{E}[\varphi_j \mid a = 0] - \mathrm{E}[\varphi_j \mid = 1]$ a $\boldsymbol{return\ f^*}$ $End\ For$

3.1 数据预处理

我们使用两个公开可用的重症监护数据集：MIMIC-IV-ED [16] 和 eICU协作研究数据库 [17]。为确保一致性并聚焦分析，性别被视为二元敏感属性 $A \in \{0,1\}$，其中 $A = 1$ 表示 男性，$A = 0$ 表示 女性。分析中排除所有性别值未定义或非二元的记录。分类特征采用标签编码，连续特征则进行归一化。每个数据集均按 80/20 的分层划分拆分为训练集和测试集，以保留敏感属性和目标类别的原始分布。形式化地，令 $X \in \mathrm{R}^{m \times d}$ 表示特征矩阵，$Y \in \{0,1\}^m$ 表示二元结果标签，$A \in \{0,1\}^m$ 表示 m 个样本的受保护属性向量。

3.2 缓解前偏差检测

首先使用二元逻辑损失训练一个标准的 XGBoost分类器 $f_\theta: X \to \hat{Y}$，在预处理后的数据上进行训练，损失函数如公式 (1) 所示，其中 $\hat{p}_i = f_\theta(x_i)$ 是预测概率，例如 i。

$$\mathcal{L}_{log} = -\frac{1}{m}\sum_{i=1}^{m}[y_i log\hat{p}_i + (1 - y_i)\log(1 - \hat{p}_i)] \quad (1)$$

3.2.1 通过 *SHAP* 进行模型解释

为诊断潜在偏见的来源和程度，我们采用 SHAP（SHapley 加性解释）[38]。对于给定的预测结果，SHAP 基于合作博弈论，为每个特征 j 赋予贡献值 ϕ_j。我们使用 TreeSHAP [44] 变体，以便对基于树的模型进行高效计算。我们生成汇总图（例如蜂群图），根据平均绝对 SHAP 值（$|\phi_j|$）对特征进行排序，从而直观呈现模型的决策逻辑，并识别可能充当性别代理变量的特征。

3.2.2 偏差量化

我们使用三个互补的公平性指标来量化基线人口统计差异：

- 统计均等差异（SPD）：使用公式（2）衡量组之间正向预测率的差异。

$$SPD = P(\hat{Y} = 1 \mid A = 1) - P(\hat{Y} = 1 \mid A = 0) \quad (2)$$

- 泰尔指数：一种用于衡量预测结果分布不平等的信息论度量，使用公式 (3) 计算

$$Theil = \sum_{i=1}^{m} \frac{\hat{p}_i}{p} \ln\left(\frac{\hat{p}_i}{p}\right), \quad where\ \ddot{p} = \frac{1}{m}\sum_{i=1}^{m} \hat{p}_i \tag{3}$$

- 瓦瑟斯坦距离 (W)：使用公式 (4) 量化两个组的预测概率分布之间的距离

$$W = \int_0^1 |F_1(t) - F_0(t)| dt \tag{4}$$

其中，F_1 和 F_0 分别是组 A = 1 和 A = 0 的累积分布函数。

3.3 公平性感知损失函数

为缓解已识别的偏差，我们构建了一个自定义损失函数，在标准预测损失中加入公平性惩罚项。总损失 L_{total} 根据公式（5）定义。

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{log} + \lambda \cdot (w_1 \cdot SPD + w_2 \cdot Theil + w_3 \cdot W) \tag{5}$$

其中，L_{log} 是二元交叉熵损失，$\lambda \in R^+$ 是用于控制整体公平性惩罚强度的正则化超参数，而w_1、w_2、$w_3 \in [0,1]$ 是用于平衡各项公平性指标（统计均等差异、泰尔指数、瓦瑟斯坦距离）贡献的权重。

3.4 缓解后训练与评估

公平性感知损失函数被集成到自定义的 XGBoost 训练循环中。我们使用贝叶斯优化结合交叉验证，对超参数 λ、w_1、w_2、w_3 进行优化，以找到预测准确率与公平性之间的最佳权衡。最终模型在留出测试集上进行评估。我们将缓解后的公平性指标（统计均等差异、泰尔指数、瓦瑟斯坦距离）与缓解前的值进行比较，以量化偏差减少情况。我们在去偏模型上重复进行 SHAP 分析，以展示特征重要性的变化以及对性别代理特征依赖程度的降低，从而为缓解过程提供可解释性洞见。

4. 实验设置、结果与讨论 n

本节介绍对 FairMed - XGB 框架的实证评估。我们详细说明实验设置，报告偏见缓解方面的定量与定性结果，并全面讨论研究发现、其影响及未来方向。

4.1 实验设置

4.1.1 数据集与队列构建

该框架在两个大规模、公开可用的重症监护数据库上进行了评估：

- MIMIC-IV-ED [16]：包含约 28 万次急诊就诊记录，时间范围为 2011—2019年，涵盖人口统计信息、生命体征、实验室结果、分诊评分、诊断和患者结局。

- eICU协作研究数据库 [17]：包含美国 200 多家机构在 2014—2015年间超过 20 万次重症监护病房住院记录，提供详细的临床测量数据、干预措施和结局。

为评估模型在不同临床情境下的表现，我们构建了七个不同的队列。来自 MIMIC-IV-ED 的队列：*diagnosis*。 *bias, medrecon. bias, triage. bias*，以及 *vitalsign*。 *bias*。来自 eICU 的队列：*patient_care*、 *patient_diagnosis* 和 *patient_treatment*。所有队列均经过筛选，仅纳入成年患者（年龄 ≥ 18）且性别标签为有效二元值的记录。分类变量采用标签编码，连续变量则进行归一化。

4.1.2 实施与评估方案

每个队列均按照性别属性进行分层抽样，划分为训练集（80%）和测试集（20%），以保持分布不变。基础分类器为 XGBoost，所有实验均在 NVIDIA RTX 3050 GPU 上进行。公平性感知损失函数（第 3.3 节）被集成到训练循环中。超参数——正则化因子 λ 和公平性指标权重 w_1, w_2, w_3——通过贝叶斯优化结合五折交叉验证进行优化，以准确率和公平性的组合目标最大化为准。

模型性能通过以下指标进行评估：

- 公平性指标：测试集上的统计均等差异（SPD）、泰尔指数和 Wasserstein 距离（W）。
- 可解释性：在缓解前后进行 SHAP（TreeSHAP）分析，以解释特征贡献。

- 预测性能：通过监测整体准确率和 AUC-ROC，评估公平性与效用权衡。

4.2 结果

4.2.1 缓解前的基线差异

标准 XGBoost 模型在所有队列中均表现出显著的性别预测差异。表 2 总结了基线公平性指标。

- MIMIC-IV-ED 队列：$medrecon.bias$ 队列的初始偏见最为严重，统计均等差异（SPD）为 0.63，泰尔指数为 ~4×104，表明预测分布高度倾斜。$diagnosis.bias$ 和$vitalsign.bias$ 队列也呈现出显著的不平衡（SPD 分别为 = 0.37 和 0.52）。

- eICU 队列：偏见现象更加突出。三个队列的 SPD 值均极高（~0.98-0.99），泰尔指数也达到 105 量级，证实基线模型存在明显的系统性偏差，倾向于某一性别群体。

这些结果从定量角度证实，在未施加公平性约束的情况下，基于不平衡临床数据训练的模型可能会延续并加剧已有的人口统计差异。

4.2.2 公平性感知缓解的效果

采用 FairMed - XGB 框架和贝叶斯优化的公平性损失，显著降低了所有数据集中的偏见。表 2 汇总并比较了缓解前后的关键公平性指标。

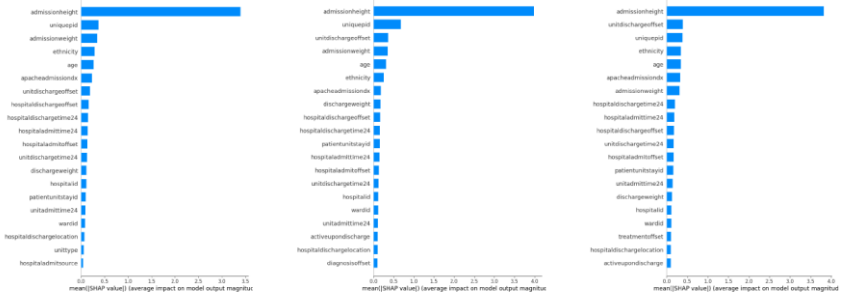
表 2。代表性队列缓解前后的公平性指标汇总

队列	女性预测%	Male预测%	SPD	泰尔	瓦瑟斯坦	女性预测%	Male预测%	SPD	泰尔	瓦瑟斯坦
	缓解前阶段					缓解前阶段				
	MIMIC-IV 急诊科									
诊断	59.36	40.63	0.37	6444.92	0.16	49.71	50.28	0.24	0.05	0.07
用药核对	62.77	37.22	0.63	40344.23	0.32	63.71	36.29	0.31	0.28	0.11
分诊	59.81	40.18	0.24	3778.09	0.12	52.97	47.02	0.20	0.05	0.071
生命体征	56.45	43.54	0.52	13589.49	0.23	52.87	47.12	0.21	0.06	0.07
	电子重症监护病房									
care	45.13	54.86	0.99	181473.06	0.95	44.86	55.13	0.80	0.53	0.65
诊断	43.54	56.45	0.98	163423.39	0.96	43.32	56.67	0.87	0.62	0.75
治疗	43.45	56.54	0.99	226915.1	0.97	43.56	56.73	0.89	0.64	0.76

- 偏差减少：缓解后，MIMIC-IV-急诊科的 SPD 降低了 40-51%，eICU 队列降低了 10-19%。泰尔指数的改善最为显著，下降了 4-5 个数量级，接近于零

（~0.06-0.65），表明已实现近乎完美的分布公平性。瓦瑟斯坦距离也降低了约 20-72%，意味着不同性别群体之间的预测分数分布重叠程度显著提高。

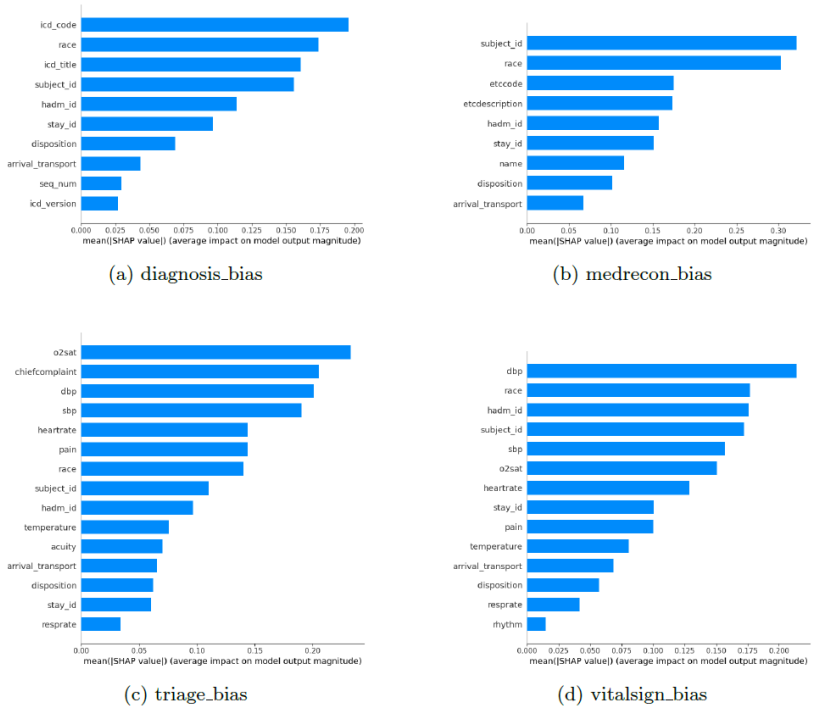
- 偏差减少：缓解后，MIMIC-IV-急诊科的统计均等差异降低了40-51%，eICU队列降低了10-19%。最显著的改善体现在泰尔指数上，其降低了4-5个数量级，接近于零（~0.06-0.65），表明已实现近乎完美的分布公平性。瓦瑟斯坦距离也减少了约20-72%，说明不同性别群体之间的预测分数分布重叠程度大幅提高。



(a) patient_care (b) patient_diagnosis (c) patient_treatment

图 1：eICU数据集中所有自定义队列的 SHAP可视化。

4.2.3 SHAP 分析得出的可解释性洞见



(a) diagnosis_bias (b) medrecon_bias

(c) triage_bias (d) vitalsign_bias

图 2：MIMIC-IV-ED 数据库中所有自定义队列的 SHAP可视化。

SHAP 分析提供了关键且可解释的证据，揭示了偏见如何表现出来并随后得到缓解（图1和图2）。

- 缓解前：基线模型的 SHAP 汇总图显示，特定临床特征和行政特征（例如特定诊断代码、年龄、分诊评分、最高心率）是最主要的贡献因素。这些特征往往是性别的强代理变量，导致预测值存在差异。

- 缓解后：应用 FairMed - XGB 后，这些代理特征的 SHAP 值分布在不同性别群体间变得显著更加均匀。特征重要性的整体排序也发生了变化，减少了模型对虚假性别相关信号的依赖，并在预测中更均衡地利用临床相关特征。

4.3 讨论

4.3.1 研究结果解读

结果验证了本研究的核心假设：在贝叶斯优化的指导下，基于多指标和惩罚项的方法能够有效缓解临床预测模型中的性别偏差。对 SPD（奇偶均等）、泰尔（分布均等）和瓦瑟斯坦距离（得分分布对齐）进行联合优化，从互补角度解决偏差问题，从而实现稳健的去偏。准确率得以保持，这证实公平性与实用性并非相互排斥，而是可以通过精心设计的目标函数进行联合优化。

SHAP可视化将缓解过程从“黑箱”转变为可审计、透明的过程。临床医生和数据管理者不仅能看到偏见是否减少，还能了解如何减少——通过观察哪些特征贡献经过了重新校准。这对于建立信任，以及在受监管的医疗环境中开展模型审计至关重要。

4.3.2 临床人工智能的影响

在标准模型中发现的普遍基线偏差，凸显了在医疗保健领域部署“开箱即用”机器学习所蕴含的重大风险。FairMed - XGB 等框架对于开发符合公平伦理原则的可信人工智能至关重要。此类框架既能提供定量公平性保证，又能提供可解释性洞见，因此能够实现：

- 法规遵从：满足医疗器械算法公平性方面不断涌现的标准。
- 临床应用：为医疗保健提供者配备易于理解的工具，增强他们对人工智能辅助决策的信心。• 健康公平：通过确保模型在不同人口统计学群体中表现公平，积极助力减少结局差异。

4.3.3 局限性与未来工作

本研究存在若干局限性，也为未来富有成效的研究指明了方向：

- 二元性别框架：本研究将性别视为二元构念。未来研究必须拓展至多类别和非二元敏感属性，并关注交叉性公平（例如，考量性别、种族和社会经济地位之间的交互作用）。• 任务泛化：尽管已在多种预测任务上进行了测试，仍有必要将 FairMed - XGB 应用于更广泛的临床结局（例如住院时长、特定并发症风险）。

- 实时部署：开发机制，在运行中的临床决策支持系统内持续或实时开展公平性监测，是实现实际部署的关键下一步。• 因果视角：更深入地整合因果推断方法，有助于区分不公平的统计关联与合理的性别临床相关因素，从而实现更细致的缓解。

5. 结论

本研究提出了 FairMed - XGB，这是一个旨在实现重症监护机器学习模型公平性的综合框架。通过将贝叶斯优化的多指标公平性惩罚项与基于 SHAP 的可解释性相结合，FairMed 在 MIMIC-IV-ED 和 eICU数据集中成功减少了基于性别的预测差异，同时保持了较高的预测准确率。该框架超越了单纯的偏差校正，提供了可操作的透明度，展示了偏差如何得到缓解。这项工作为在高风险医疗场景中部署更公平、可信的机器学习模型提供了一条务实、合乎伦理且透明的路径，并为医学领域算法公平性这一更广泛的目标作出了贡献。

参考文献

1. Olalekan Kehinde, A.（2025年）。运用机器学习构建医疗保健预测模型，以改善患者结局管理。*Int Res J Mod Eng Technol Sci*，*7*(1)，1465。2. Ghimire, P., & Adhikari, S.（2025年）。通过数据分析推进智慧医疗：机器学习与预测模型在临床决策中的整合研究。计算系统、算法与新兴技术进展，*10*(4)，21-36。3. Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., ... & Das, R.（2016年）。利用最少的电子健康记录数据预测重症监护病房中的脓毒症：一种机器学习方法。*JMIR medical informatics*，*4*(3)，e5909。4. Awwad, Y., Fletcher, R., Frey, D., Gandhi, A., Najafian, M., & Teodorescu, M.（2020年）。探索国际发展领域机器学习的公平性。CITE MIT D-Lab。5. Hoffmann, D. E., & Goodman, K. E.（2022）。公共卫生紧急事件期间稀缺医疗资源的分配：我们能否考虑生理性别？*Hous. J. Health L. & Pol'y*，*22*，175。6. Qi, M.（2022）。运用机器学习公平性、人工智能和优化解决健康不公平问题。伦斯勒理工学院。7. Andrich, A.（2025年）。检测政治传播中性别偏差的计算与纵向研究方法（博士论文，论文，伊尔梅瑙，伊尔梅瑙工业大学，2025年）。8. Long, E. F., & Mathews, K. S.（2018年）。滞留患者：重症监护病房和医院占床率激增对患者流转的影响。*Production and operations management*，*27*(12)，2122-2143。

9. Mohammed, S., Matos, J., Doutreligne, M., Celi, L. A., & Struja, T.（2023年）。脓毒症患者侵入性重症监护治疗中的种族差异：基于MIMIC-IV高分辨率电子健康记录的分析。《耶鲁生物学与医学杂志》，96(3)，293。10. O'Brien, N., Van Dael, J., Clarke, J., Gardner, C., O'Shaughnessy, J., Darzi, A., & Ghafur, S.（2022）。解决数据驱动的健康技术中的种族和族裔不平等。11. Lim, J., Kim, Y., Kim, B., Ahn, C., S hin, J., Yang, E., & Han, S.（2023年）。BiasAdv：用于模型去偏的偏差对抗式增强。载于IEEE/CVF计算机视觉与模式识别会议论文集（第3832-3841页）。12. Zhang, Y., & Sang, J.（2020年10月）。迈向准确率—公平性悖论：用于视觉去偏的基于对抗样本的数据增强。载于第28届ACM多媒体国际会议论文集（第4346-4354页）。13. Chen, R. J., Wang, J. J., Williamson, D. F., Chen, T. Y., Lipkova, J., Lu, M. Y., ... & Mahmood, F.（2023年）。医学与医疗保健人工智能中的算法公平性。《自然-生物医学工程》，7(6)，719-742。14. Moreno-Sánchez, P. A., Del Ser, J., van Gils, M., & Hernesniemi, J.（2025年）。在医疗保健领域落实可信人工智能的设计框架：临床应用的要求、权衡与挑战。arXiv预印本 arXiv:2504.19179。15. Greengrass, C. J.（2025年）。变革临床推理——人工智能在支持人类认知局限方面的作用。《数字健康前沿》，7，1715440。16. Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., & Horng IV, S.（2021年）。MIMIC-IV-ED。PhysioNet。17. Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O.（2018年）。eICU协作研究数据库：一个免费开放的多中心重症监护研究数据库。《科学数据》，5(1)，1-13。18. Fain, J. A.（2017年）。NHANES：免费公共数据集的使用。《糖尿病教育者》，43(2)，151-151。19. Allen, N. E., Sudlow, C., Peakman, T., Collins, R., & 英国生物样本库。（2014年）。英国生物样本库数据：欢迎获取。《科学转化医学》，6(224)，224ed4-224ed4。20. Biobank, Mayo Blegen Ashley L. 18 Wirkus Samantha J. 18 Wagner Victoria A. 18 Mey-er Jeffrey G. 18 Cicek Mine S. 10 18, & All of Us研究示范项目团队 Choi Seung Hoan 14 http://orcid. org/0000-0002-0322-8970 Wang Xin 14 http://orcid. org/0000-0001-6042-4487 Rosenthal Elisabeth A. 15。（2024年）。All of Us研究计划中的基因组数据。《自然》，627(8003)，340-346。21. Andersson, C., Johnson, A. D., Benjamin, E. J., Levy, D., & Vasan, R. S.（2019年）。弗雷明汉心脏研究的70年传承。《自然综述：心脏病学》，16(11)，687-698。22. Iachan, R., Pierannunzi, C., Healey, K., Greenlund, K. J., & Town, M.（2016年）。行为风险因素监测系统（BRFSS）数据的全国加权。《BMC医学研究方法学》，16(1)，155。23. Taub, N.（1995年）。差异影响分析在追求性别平等中的相关性。《Seton Hall宪法法评论》，6，941。24. Liang, X. S.（2021年）。归一化多变量时间序列因果分析与因果图重构。《熵》，23(6)，679。25. Zhang, B. H., Lemoine, B., & Mitchell, M.（2018年12月）。通过对抗学习减轻不良偏差。载于2018年AAAI/ACM人工智能、伦理与社会会议论文集（第335-340页）。26. Kamalaruban, P., Pi, Y., Burrell, S., Drage, E., Skalski, P., Wong, J., & Sutton, D.（2024年11月）。评估交易欺诈模型中的公平性：公平性指标、偏差审计，

及挑战。载于《第5届ACM人工智能金融国际会议论文集》（第555-563页）。27. Veitch, V., D'Amour, A., Yadlowsky, S., & Eisenstein, J. (2021). 对虚假相关性的反事实不变性：为何以及如何通过压力测试。arXiv预印本，arXiv:2106.00545。28. Yazdi, F., & Asadi, S. (2025年). 增强心血管疾病诊断：优化集成学习的力量。《IEEE Access》，13，46747-46762。29. Meng, C., Trinh, L., Xu, N., Enouen, J., & Liu, Y. (2022). MIMIC-IV数据集上深度学习模型的可解释性与公平性评估。《Scientific Reports》，12(1)，7166。30. Mohammed, S., Matos, J., Doutreligne, M., Celi, L. A., & Struja, T. (2023年). 脓毒症患者侵入性重症监护治疗中的种族差异：基于MIMIC-IV高分辨率电子健康记录的分析。《耶鲁生物学与医学杂志》，96(3)，293。31. Miller, S. (2023年). 使用数据平衡算法纠正心血管数据集中女性患者代表不足的问题。32. Yang, J., Soltan, A. A., Eyre, D. W., Yang, Y., & Clifton, D. A. (2023年). 用于缓解临床机器学习中算法偏见的对抗训练框架。《NPJ数字医学》，6(1)，55。33. Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (2023年). 机器学习模型中的偏见与不公平性：关于数据集、工具、公平性指标以及识别和缓解方法的系统综述。《大数据与认知计算》，7(1)，15。34. Mures, O. A., Taibo, J., Padrón, E. J., & Iglesias-Guitian, J. A. (2025年). 通过基于神经网络的合成数据生成缓解表格数据中的类别不平衡：全面综述与软件库。Authorea预印本。35. Mulomba, C. M., Kiketa, V. M., Kutangila, D. M., Mampuya, P. H., Mukenze, N., Kasunzi, L. M., ... & Kasereka, S. K. (2025年). 将因果机器学习应用于时空数据分析：机遇与挑战研究。《IEEE Access》。36. Hüser, M. (2021年). 重症监护病房患者监测的机器学习方法（博士论文，苏黎世联邦理工学院）。37. Liu, Z., Sarkani, S., & Mazzuchi, T. 公平性优化动态聚合（Foda）：一种在异构环境中实现公平联邦学习的新方法。可通过SSRN 5145003获取。38. Lundberg, S. M., & Lee, S. I. (2017年). 解释模型预测的统一方法。《神经信息处理系统进展》，30。39. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016年8月). “我为什么应该相信你？”解释任意分类器的预测。载于《第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集》（第1135-1144页）。40. Adebayo, J. A. (2016年). FairML：用于诊断预测建模中偏见的工具箱（博士论文，麻省理工学院）。41. Jo, N., Aghaei, S., Benson, J., Gomez, A., & Vayanos, P. (2023年8月). 学习最优公平决策树：可解释性、公平性与准确率之间的权衡。载于《2023年AAAI/ACM人工智能、伦理与社会会议论文集》（第181-192页）。42. Mothilal, R. K., Sharma, A., & Tan, C. (2020年1月). 通过多样化的反事实解释说明机器学习分类器。载于《2020年公平、问责与透明度会议论文集》（第607-617页）。43. Abualrous, R., Zouzou, H., Zgheib, R., Hasan, A., Hijazi, B., & Kermani, A. (2025年). 公平性感知智能强化（FAIR）：人工智能驱动的医院排班框架。《Information》，16(12)，1039。

44. Bifet, A.、Read, J. 和 Xu, C.（2022）。线性树 SHAP。神经信息处理系统进展，*35*，25818-25828。