Investigating Bias in AI for Sepsis Diagnosis: Analyzing Ethnic Disparities in ICU Decision-Making

Armand Patel, Joshua Payapulli, Paul Yoo
University of Southern California, Los Angeles, CA, USA
(Dated: May 25, 2025)

Artificial intelligence (AI) is increasingly used to support clinical decision-making, including early detection of sepsis. However, these models risk amplifying existing disparities, particularly across racial and ethnic groups. Using the MIMIC-III database, we evaluate three models; logistic regression, random forest, and the multi-layer perceptron (MLP) to assess subgroup-level performance disparities in sepsis prediction. We analyze error rates across ethnic groups, interpret model behavior using SHAP values, and apply threshold adjustment to mitigate bias. Our findings underscore the need for fairness-aware evaluations in clinical AI.

Keywords: Bias, Healthcare AI, Sepsis, MIMIC-III, ICU Ethnic Disparities

I. INTRODUCTION

Artificial intelligence (AI) models are becoming integral to clinical workflows, particularly for high-stakes tasks like sepsis diagnosis. While these tools promise gains in accuracy and efficiency, they may also encode and reinforce structural inequities present in the underlying data. Recent studies have highlighted that algorithmic bias can emerge throughout the machine learning pipeline; from data collection to model deployment, often disadvantaging marginalized populations [1]. Despite this, fairness evaluations remain unexplored in clinical AI research [2], limiting our understanding of how models perform across demographic subgroups.

In this work, we investigate how model complexity and fairness interventions affect subgroup disparities in sepsis prediction. Using the MIMIC-III database, we train and evaluate logistic regression, random forest, and MLP models across self-reported ethnic groups. We examine groupwise error rates, use SHAP values for interpretability, and apply adversarial debiasing and post-hoc threshold adjustment to mitigate bias.

We aim to answer two central questions: (1) How does model complexity, ranging from logistic regression to deep learning, affect both predictive performance and fairness in sepsis diagnosis? (2) How effective are bias mitigation strategies in reducing ethnic disparities in prediction?

II. RELATED WORKS

Bias in artificial intelligence (AI) systems within healthcare has become an area of increasing concern, especially in relation to racial and ethnic disparities. Several studies have demonstrated that predictive models used in clinical decision making often replicate structural inequities present in the data they are trained on. Chesley et al. (2023), for example, found that Black patients hospitalized with sepsis experienced significantly longer lengths of stay than White patients, even when controlling for clinical severity and hospital capacity [3]. These findings suggest that disparities in outcomes may stem from systemic biases in care delivery that are not fully accounted for by clinical factors.

A. Bias in AI-Driven Sepsis Diagnosis

Sepsis prediction models in particular have been scrutinized for uneven performance across demographic groups. Komorowski et al. (2018) developed one of the earliest AI-based clinical decision support systems for sepsis, training it on historical ICU data [4]. While their model demonstrated improved treatment recommendations, it also raised concerns about the risk of reproducing biases encoded in past medical decisions. Fleuren et al. (2020), in a systematic review of sepsis prediction models, found that most models were evaluated using retrospective data with little subgroup analysis, noting reduced accuracy in racial minority populations due to class imbalance and biased feature selection [2].

Obermeyer et al. (2019) further highlighted these issues in a broader context, showing that widely used healthcare risk algorithms underestimated the needs of Black patients, resulting in under-allocation of care [5]. Their work emphasized the need for subgroup-level evaluation in clinical AI systems to prevent the reinforcement of existing disparities. Similarly, Cross et al. (2024) examined ICU-based AI tools and found that even models with high overall accuracy often misclassified patients from marginalized groups at higher rates [1]. Their findings support the need to assess fairness not just at the model level but also at the level of subgroup outcomes.

More broadly, researchers have called for better frameworks to guide the development and evaluation of AI models in healthcare. Rajkomar et al. (2019) and Gianfrancesco et al. (2018) have both pointed to the risks of relying on biased electronic health record (EHR) data, arguing that performance gains in predictive tasks can come at the cost of equitable treatment unless fairness is explicitly considered [6].

B. Disparities in Sepsis Outcomes

Disparities in sepsis-related care extend beyond prediction to treatment and outcomes. DiMeglio et al. (2018) reported that Black and Hispanic patients face higher sepsis-related mortality rates than their White counterparts, even under standardized treatment protocols [7]. These disparities arise from a mix of patient-level factors, such as higher rates of comorbidities, and systemic issues like limited access to quality care. Importantly, minority patients are more likely to receive treatment in under-resourced hospitals with lower adherence to clinical guidelines, which further contributes to worse outcomes.

C. Fairness Gaps in Existing Work

Despite the recognition of these disparities, most sepsis prediction research has focused on overall model performance rather than fairness. Fleuren et al. (2020) noted that among over 100 sepsis models reviewed, few assessed groupwise performance metrics, and even fewer implemented fairness interventions. Bias mitigation strategies, such as adversarial debiasing and threshold adjustment, have been proposed in theoretical work but are rarely evaluated in empirical studies specific to clinical AI. As a result, the literature lacks clarity on whether these approaches can meaningfully reduce disparities without compromising diagnostic utility.

Our work contributes to this gap by focusing explicitly on ethnic disparities in sepsis prediction. We evaluate how model complexity; from logistic regression to deep learning, affects both predictive accuracy and fairness across subgroups. We further assess whether post hoc mitigation strategies can reduce error disparities between ethnic groups, providing evidence to guide the development of more equitable clinical AI systems.

III. METHODOLOGY

A. Data Source and Cohort Construction

We use the MIMIC-III (Medical Information Mart for Intensive Care III) database, which contains over 53,000 ICU admissions, offering a demographically diverse patient cohort. Sepsis-positive cases were identified using ICD-9 codes (99591, 99592, 78552) consistent with prior clinical studies. We limited our analysis to patients with non-missing values for age, gender, ethnicity, and insurance type, yielding two cohorts: sepsis-positive patients and a control group of sepsis-negative ICU patients. These were combined into a binary classification dataset.

B. Feature Selection and Preprocessing

We selected age, gender, self-reported ethnicity, and insurance type as features of interest due to their clinical and fairness relevance. Categorical features were one-hot encoded. Because the original dataset was highly imbalanced, we applied random oversampling of the minority class to ensure class balance. The data was split into training and testing sets using an 80/20 stratified split on the outcome variable.

C. Modeling and Evaluation

We trained three predictive models: a logistic regression classifier, a random forest, and a feedforward multilayer perceptron (MLP). The MLP consisted of two hidden layers (128 and 64 units, ReLU activation) and was trained using the Adam optimizer with early stopping. All models were evaluated using standard classification metrics: AUROC, precision, recall, F1-score, and confusion matrices.

D. Explainability and Fairness Analysis

We applied SHAP (SHapley Additive exPlanations) to interpret model predictions and identify influential features. Across models, demographic and socioeconomic variables; particularly ethnicity and insurance type frequently ranked among the top contributors, raising concerns about potential bias. To evaluate group-level fairness, we computed the false negative rate (FNR) and false positive rate (FPR) across self-reported ethnic groups. These metrics capture underdiagnosis and overdiagnosis, respectively.

IV. RESULTS

Table I compares the performance of three classification models used for sepsis prediction: logistic regression, random forest, and a multi-layer perceptron (MLP). Among these, the random forest model achieved the strongest results overall, with an AUROC of 0.83 and the highest F1-score and precision, outperforming both the logistic regression and MLP baselines.

The random forest offered a strong balance between sensitivity and specificity, reflected in its F1-score of 0.750 and precision of 0.740. By contrast, logistic regression produced more conservative predictions, leading to lower precision and weaker overall discriminative ability. The MLP classifier, despite its additional complexity, did not yield meaningful improvements. Its recall was slightly better than that of logistic regression, but precision and F1-score remained lower than those of the random forest. These results suggest that simply increasing model complexity does not guarantee better perfor-

mance, particularly when the input features are limited to demographic and administrative data. Without richer clinical signals, the MLP may have picked up on spurious patterns, resulting in small gains that do not justify the added complexity.

Model	AUROC	Prec.	Rec.	F1
Logistic Regression	0.695	0.595	0.630	0.610
Random Forest	0.830	0.740	0.770	0.750
MLPClassifier	0.710	0.582	0.610	0.600

TABLE I. Performance comparison of logistic regression, random forest, and multi-layer perceptron (MLP) classifiers on sepsis prediction.

A. Statistical Testing

We found a significant association between ethnicity and sepsis prevalence in the dataset ($\chi^2 = 222.38$, $p < 0.001$), suggesting that baseline disparities exist across demographic groups. These differences may contribute to biased model behavior by reflecting structural inequities in the data. In contrast, a one-way ANOVA on a proxy for time-to-diagnosis showed no significant variation across the five largest ethnic groups ($F = 0.30$, $p = 0.879$), indicating no evidence of disparity in that particular outcome.

B. Logistic Regression Analysis

The logistic regression model exhibited notable disparities in predictive performance, particularly in its sensitivity to sepsis-positive cases. Despite identifying many true positives, the model frequently underdiagnosed positive cases, raising concerns about its clinical reliability. In high-stakes healthcare settings, such underdiagnoses can delay treatment and increase the risk of adverse outcomes.

To better understand the model's decision-making, we used SHAP (SHapley Additive exPlanations) to analyze feature importance. As shown in Fig. 1, the model's predictions were strongly influenced by demographic and socioeconomic attributes such as `insurance_Medicare`, `insurance_Medicaid`, `age`, and several `ethnicity` categories. The prominence of these non-clinical features suggests that the model may be capturing patterns in the data distribution that reflect existing societal or systemic biases.

We further evaluated fairness using group-wise False Negative and False Positive Rates. These metrics varied substantially across ethnic groups, indicating potential disparities in how the model performs for different subpopulations. A Chi-square test confirmed a statistically significant association between ethnicity and prediction outcome ($\chi^2 = 222.38$, $p < 0.001$), highlighting the uneven error distribution.

These findings underscore that despite a model being interpretable and transparent it does not mean it is inherently fair. Biases present in the training data can propagate through the model and lead to inequitable outcomes. Mitigating such effects requires deliberate intervention through fairness-aware training methods.

FIG. 1. SHAP summary plot for logistic regression. Top features include age, insurance type, and ethnicity.

C. Random Forest Classifier Analysis

To evaluate whether model complexity improves both prediction and fairness, we trained a random forest with 50 estimators and a maximum depth of 8. The model outperformed logistic regression, achieving an AUROC of 0.83, along with improved precision (0.74), recall (0.77), and F1-score (0.75).

To understand which variables influenced predictions, we used SHAP. As shown in Fig. 2, the most influential features were largely demographic: `age`, insurance categories (e.g., `Medicare`, `Medicaid`), and `ethnicity`. These findings suggest that the model prioritized structural and social factors present in the training data rather than physiological variables.

To further explore how these variables interact, we generated a SHAP interaction plot focused on `age` (Fig. 2). This plot visualizes how the contribution of age to the prediction depends on its interaction with other features. For instance, age interacts differently with insurance types across the population: among older pa-

tients (shown in red), certain insurance categories magnify the model's predicted risk, while for younger individuals (blue), those same categories may dampen it. The split between the upper and lower bands in the plot represents the two interacting features; `age` and the secondary variable it modifies.

Although the random forest improved overall predictive performance, disparities in error rates across ethnic groups persisted. Some subpopulations continued to experience higher false positive or false negative rates, indicating that increased flexibility alone does not ensure fairness. Instead, the model may be more precisely capturing existing biases encoded in the data.

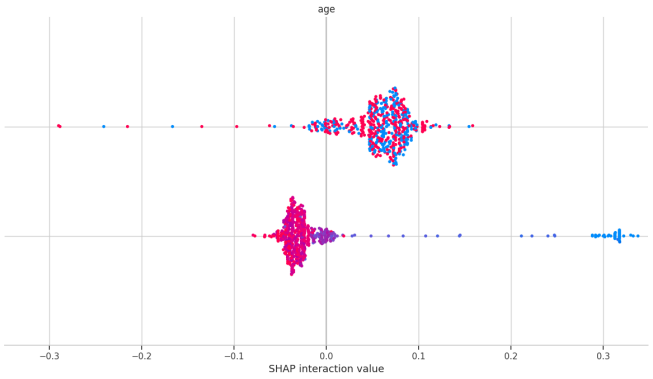

FIG. 2. SHAP summary plot for the random forest model. Feature importance values reflect each variable's average marginal contribution to model predictions.

D. Deep Learning Classifier Analysis

We trained a feedforward Multi-Layer Perceptron (MLP) with two hidden layers (128 and 64 units), using ReLU activations, Adam optimizer, and early stopping. Input features included standardized, one-hot encoded demographic and administrative variables. Class imbalance was addressed through stratified sampling and oversampling of the minority class.

As shown in Table I, the MLP achieved an AUROC of 0.71 and an F1-score of 0.60; roughly equal to the logistic regression and notably worse than the random forest across all metrics.

Fairness analysis revealed continued disparities in error rates across ethnic groups (Fig. 6). Despite the model's capacity for non-linear learning, it failed to improve fairness. Predictions remained highly sensitive to demographic inputs such as ethnicity and insurance type, and no substantial mitigation of group-level error gaps was observed.

Overall, the MLP neither improved predictive performance nor reduced bias. These results suggest that increasing model complexity without introducing richer clinical signals or fairness-aware training does little to address disparities for sepsis diagnosis.

E. Bias Mitigation

To improve group-level fairness, we applied two mitigation strategies to the random forest model: adversarial debiasing and post-hoc threshold adjustment.

Adversarial debiasing trains a predictor alongside an adversary that attempts to infer protected attributes from the model's internal representations. The classifier is optimized to predict sepsis while minimizing the adversary's success, encouraging group-invariant representations. While this reduced false positive rate (FPR) disparities, false negative rates (FNR) drastically increased limiting its ability to accurately diagnose sepsis.

As an alternative, we implemented threshold adjustment. Rather than removing protected attribute signals, this method retained them and adjusted classification thresholds for each ethnic group, prioritizing predictive parity over entirely removing protected attributes. We were also able to prioritize minimizing FNR rates using this technique, which further improves its clinical effectiveness as diagnosing true sepsis cases is more important than missing false alarms.

We quantified fairness using Equalized Odds gaps, the absolute differences in FNR and FPR across ethnic groups. As shown in Fig. 3, threshold adjustment yielded the lowest FNR gap (0.05), while adversarial debiasing achieved the lowest FPR gap (0.13). Both approaches improved upon the baseline models in terms of fairness, but threshold adjustment was the only technique to improve fairness without compromising accuracy.

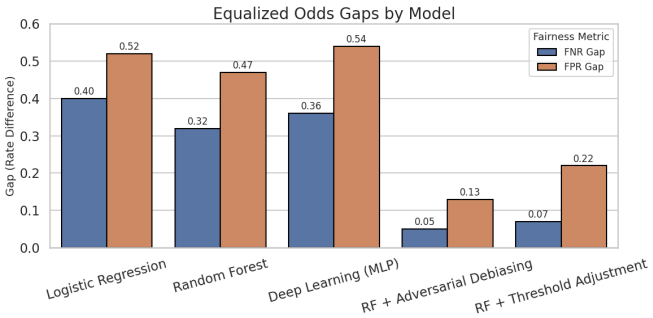

FIG. 3. Equalized Odds gaps (FNR and FPR) across models. Lower values indicate reduced disparity in misclassification rates by ethnicity.

Model	AUROC	Prec.	Rec.	F1
RF + Adv. Debiasing	0.643	0.638	0.156	0.251
RF + Thresholding	0.669	0.639	0.990	0.777

TABLE II. Performance of fairness-enhancing methods applied to the Random Forest model.

V. CONCLUSIONS

Our findings highlight the tension between predictive performance and fairness in clinical machine learning. While both logistic regression and random forest achieved reasonable accuracy, they consistently exhibited disparities in false negative and false positive rates across racial and ethnic groups. These disparities persisted as we moved from a simple linear model to a more complex ensemble classifier and also when using the multi-layer perceptron.

SHAP analyses showed that non-clinical features such as ethnicity and insurance type had outsized influence on predictions. Although these variables may correlate with care patterns or access to treatment, their prominence suggests the models may be learning from structural inequities rather than underlying physiology. The random forest model, in particular, captured more complex feature interactions, but this did not necessarily increase fairness substantially.

Despite improvements in overall performance, fairness gaps remained. Black, Hispanic, and Middle Eastern patients experienced higher false negative rates, while several Asian subgroups saw elevated false positive rates. These patterns risk reinforcing existing disparities in diagnosis and treatment, especially in critical care settings where early detection is essential.

Statistical tests confirmed that both sepsis outcomes and predicted labels varied significantly by ethnicity, suggesting these disparities are systematic. Notably, these associations persisted even after adjusting for model complexity and resampling methods.

Overall, the results make clear that accuracy alone is not enough. Without deliberate fairness interventions, models will reflect and potentially exacerbate inequities present in the data. Future work should try to integrate fairness objectives directly into model development, rather than treating them as post-hoc adjustments.

GENAI USAGE STATEMENT

ChatGPT was used solely to review grammar and improve sentence clarity throughout the writing process. No text was generated directly for inclusion in the report.

BIBLIOGRAPHY

[1] J. L. Cross, M. A. Choma, and J. A. Onofrey, Bias in medical ai: Implications for clinical decision-making, PLOS Digital Health **3**, e0000651 (2024).

[2] L. M. Fleuren, T. L. T. Klausch, J. Zwager, L. J. Schoonmade, P. W. B. Nanayakkara, A. Abu-Hanna, L. M. M. Peelen, and E. van der Veen, Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy, Intensive Care Medicine **46**, 383 (2020).

[3] C. F. Chesley, M. Chowdhury, D. S. Small, D. Schaubel, V. X. Liu, M. B. Lane-Fall, S. D. Halpern, and G. L. Anesi, Racial disparities in length of stay among severely ill patients presenting with sepsis and acute respiratory failure, JAMA **329**, 987 (2023).

[4] M. Komorowski, L. A. Celi, O. Badawi, A. E. Gordon, and A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature Medicine **24**, 1716 (2018).

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science **366**, 447 (2019).

[6] A. Rajkomar, J. Dean, and I. S. Kohane, Machine learning in medicine, New England Journal of Medicine **380**, 1347 (2019).

[7] M. DiMeglio, J. Dubensky, S. Schadt, R. Potdar, and K. Laudanski, Factors underlying racial disparities in sepsis management, Healthcare **6**, 133 (2018).

APPENDIX

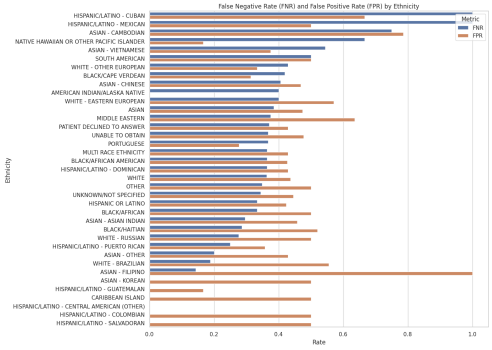

FIG. 4. False Negative Rate (FNR) and False Positive Rate (FPR) by ethnicity for the logistic regression model.

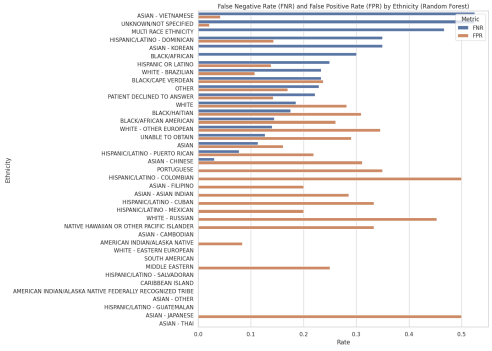

FIG. 5. False Negative Rate (FNR) and False Positive Rate (FPR) by ethnicity for the random forest model.

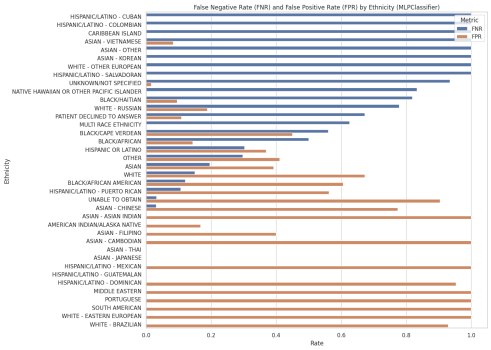

FIG. 6. False Negative Rate (FNR) and False Positive Rate (FPR) by ethnicity for the MLP classifier.

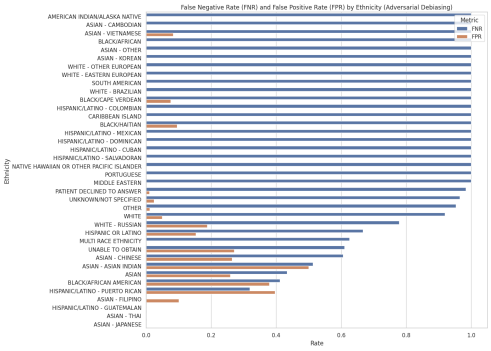

FIG. 7. False Negative Rate (FNR) and False Positive Rate (FPR) by ethnicity after adversarial debiasing.

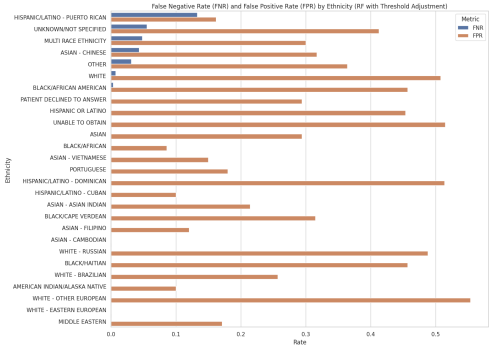

FIG. 8. False Negative Rate (FNR) and False Positive Rate (FPR) by ethnicity after threshold adjustment.

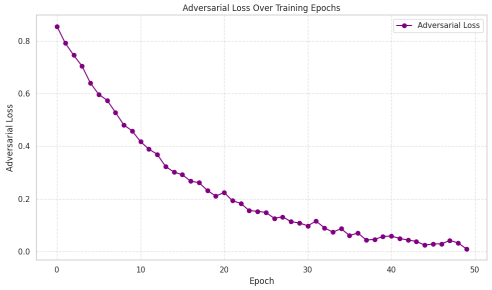

FIG. 9. Adversarial loss over training epochs.