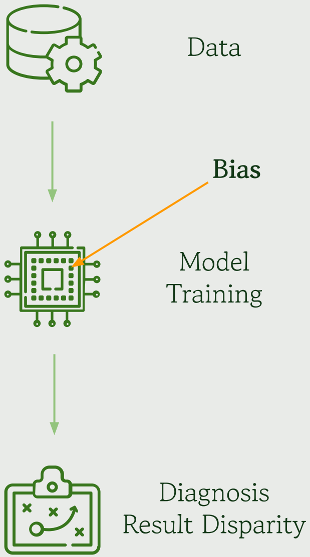
Investigating Bias in AI for Sepsis Diagnosis

Analyzing Disparities in ICU Decision-Making

Team 3 -

Joshua Payapulli
Armand Patel
Paul Yoo

Problem Description and Motivation



Rising Artificial Intelligence (AI) models for the sepsis diagnosis may **encode and perpetuate racial and socioeconomic biases, leading to disparities in patient outcomes.** Bias can originate at any stage of the development pipeline, and such bias in medical diagnosis may have critical consequences for patients.

The study aims to **uncover hidden biases using causal inference and feature attribution techniques, and to evaluate debiasing strategies** across model types.

J. L. Cross, M. A. Choma, and J. A. Onofrey, "Bias in medical AI: Implications for clinical decision-making," *PLOS Digital Health*, vol. 3, no. 11, p. e0000651, 2024. [Online]. Available: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000651

Methods

Chi Square / ANOVA

Statistical Testing and Bias Detection

Applied Chi-square tests and ANOVA to evaluate whether disparities in sepsis incidence and model predictions were statistically significant across demographic groups (e.g., race, insurance); Relationship between patient ethnicity and true sepsis incidence produced significant result ($\chi2 = 222.38$, $p < 0.001$)

SHAP

Feature Attribution

Used SHAP (SHapley Additive exPlanations) to interpret model predictions and uncover the extent to which socio-demographic features like ethnicity and insurance influenced sepsis classification.

FNR / FPR

Fairness Metrics and Disparity Evaluation

Measured False Negative Rate (FNR) and False Positive Rate (FPR) across ethnic and socioeconomic groups to quantify disparities in underdiagnosis and overdiagnosis, highlighting group-level biases.

LR vs. RF

Comparative Modeling

Compared Logistic Regression (LR) and Random Forest (RF) models to investigate how model complexity impacts predictive accuracy and fairness, analyzing trade-offs between bias and overall performance.

Debiasing

Bias Mitigation Strategy

Applied adversarial debiasing & threshold adjustment to reduce bias and added threshold adjustment to further lower false negatives across demographic groups, while maintaining model performance.

Results

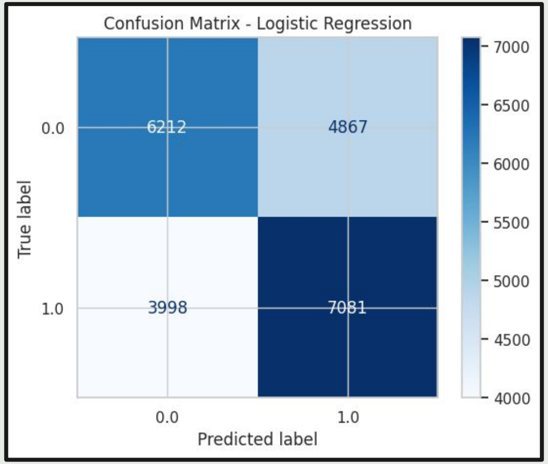
Bias Detection and Baseline Analysis

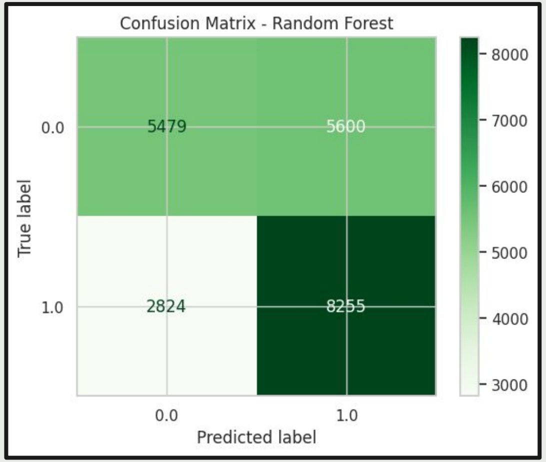
- **Black, Puerto Rican, and Middle Eastern patients** had higher FNRs — indicating underdiagnosis.
- Certain **Asian subgroups** had elevated FPRs — indicating overdiagnosis.
- SHAP analysis revealed **high influence of non-clinical features (ethnicity, insurance type)** — revealing demographic bias

Logistic Regression < **Random Forest**

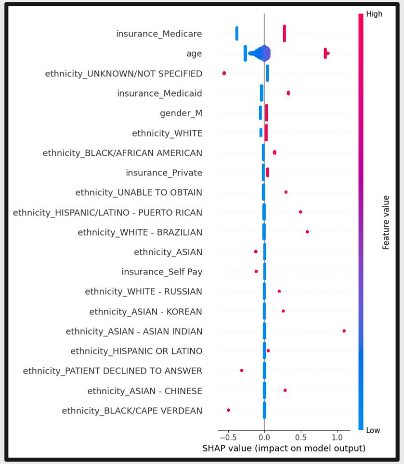
Model	AUROC	Precision	Recall	F1-Score
Logistic Regression	0.78	0.70	0.75	0.72
Random Forest	0.83	0.74	0.77	0.75

[Table 1] Logistic Regression and Random Forest Comparison: AUROC, Precision, Recall and F1-Score





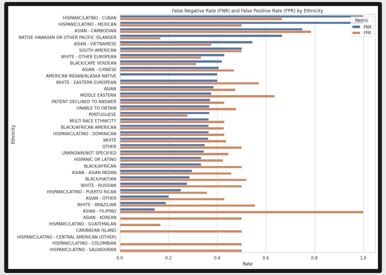
[Figure 1, 2] Confusion Matrix: Logistic Regression and Random Forest

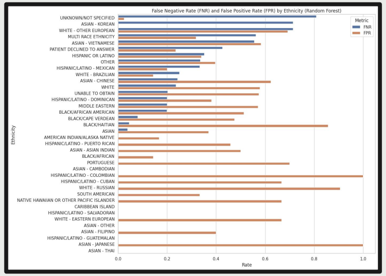


[Figure 3] SHAP Value: Logistic Regression

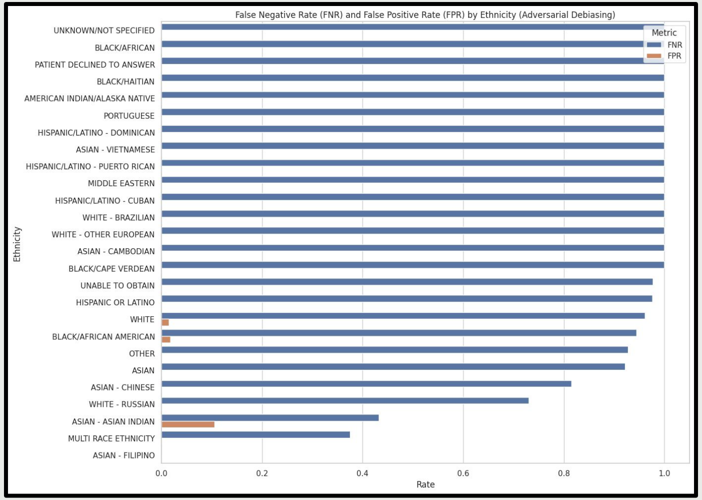
Bias Mitigation Outcomes

- Adversarial debiasing reduced bias, but **disparities in false negatives (FNR) across groups remained**.
- **Threshold adjustment further reduced FNR**, improving fairness without major performance loss.
- **Combined strategy** highlights the importance of fairness-aware calibration in clinical AI.

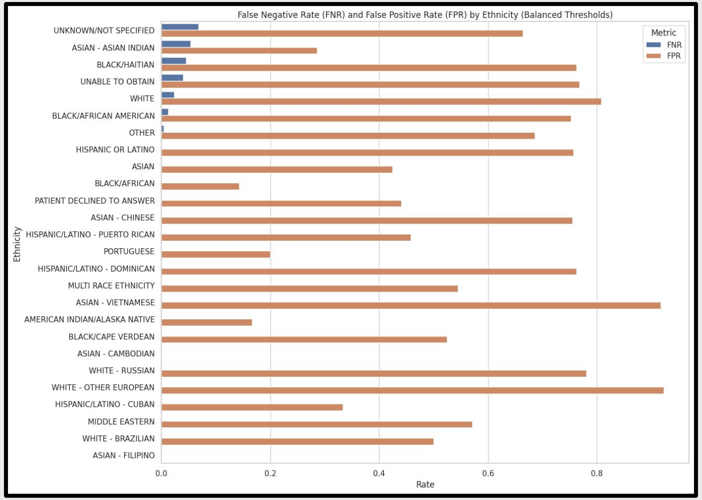




[Figure 4, 5] FNR and FPR Rate by Ethnicity before Adversarial Debiasing (LR and RF)



[Figure 6] FNR and FPR Rate by Ethnicity after Adversarial Debiasing



[Figure 7] FNR and FPR Rate by Ethnicity after Threshold Adjustment

Thank You!

- Q & A