SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Classifying, Fast and Slow: Adversarial Training for Bias Mitigation in Medical Imaging

Felipe Matsuoka, Faculdade de Ciências Médicas da Santa Casa de São Paulo; Eduardo Farina, MD;
Felipe Kitamura, MD, PhD, MS

Introduction/Background

Ethnicity bias in deep learning models poses significant ethical concerns. Leveraging Daniel Kahneman's dual-process theory in "Thinking, Fast and Slow," which distinguishes between rapid, intuitive System 1 and deliberate, analytical System 2 thinking, we propose an approach to mitigate bias in chest X-ray classification. Previous studies have demonstrated the effectiveness of adversarial methods in reducing bias, such as COVID-19 classification from electronic medical records (Zhang et al., 2021), making this approach both relevant and innovative in medical imaging.

Methods/Intervention

Our methodology employs two complementary models: a predictor model and an adversarial model. The predictor model, akin to System 1, efficiently classifies chest X-rays, identifying normal and abnormal cases. Simultaneously, the adversarial model, similar to System 2, challenges the predictions to reduce bias. We used the CheXpert dataset (Irvin et al., 2019), ensuring a balanced representation of ethnic groups through binary label adjustment and sampling techniques. During training, the adversarial model increases its error in predicting patient ethnicity, forcing the predictor model to focus on unbiased features.

Results/Outcome

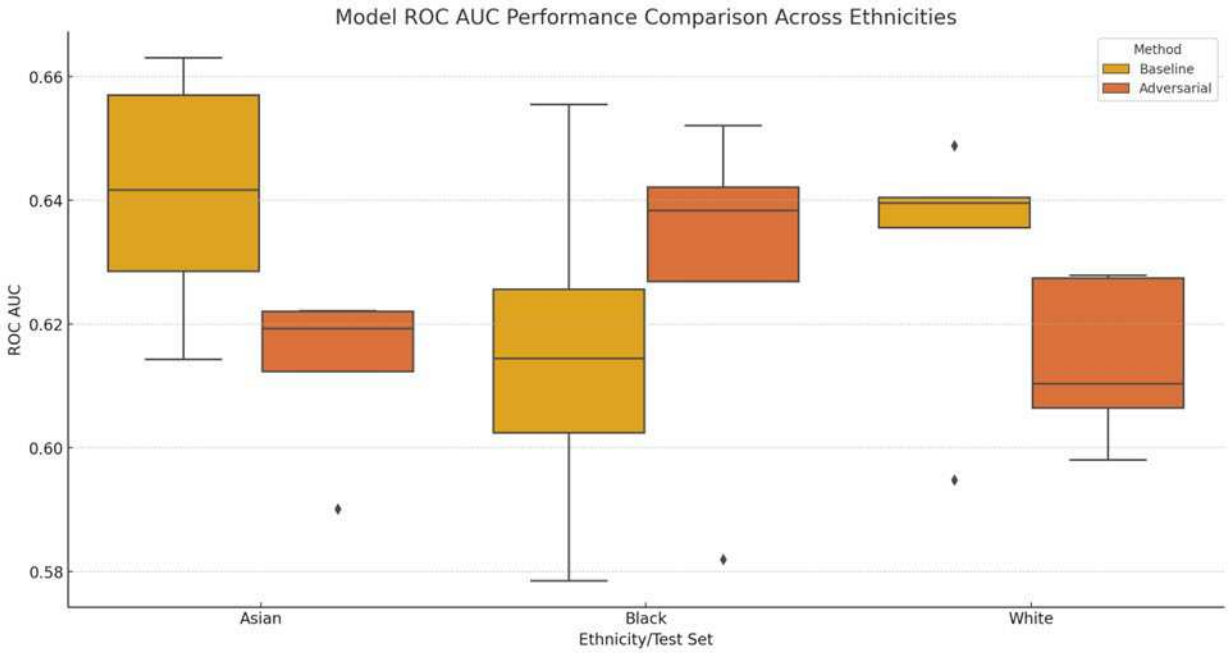
Using One-way ANOVA, we assessed the ROCAUC performance across different ethnicities for both models. The baseline model showed no significant differences across ethnicities (p-value = 0.258). Similarly, the adversarial model also exhibited no significant differences (p-value = 0.405). These findings suggest that the adversarial model maintained consistent performance across ethnic groups without introducing additional bias, highlighting the complexity of addressing ethnicity bias in medical imaging (Figure1).

Conclusion

The adversarial training framework in chest X-ray classification demonstrates an innovative approach to mitigating ethnicity bias. Despite not showing significant performance differences, this study emphasizes the importance of developing methods to address bias in medical imaging. The results suggest that achieving equitable performance across all ethnic groups is challenging, and a potential alternative could involve optimizing models for specific ethnicities.

Statement of Impact

This study introduces a novel adversarial training framework to mitigate ethnicity bias in deep learning models for medical imaging. The ANOVA results indicate that adversarial methods can maintain equitable performance across different ethnic groups. By applying principles from psychology, this research connects theoretical concepts with practical applications, advancing the development of more reliable AI systems in medical diagnostics.



ROC AUC Performance Comparison Across Ethnicities for Baseline and Adversarial Models

Keywords

Ethnicity Bias; Deep Learning; Medical Imaging; Adversarial Training; Bias Mitigation; Artificial Intelligence