Evaluating Fairness in ICU Mortality Prediction with MIMIC-IV (Literature Review)

Liujia Yu, Yun-Jing Chang

(Dated: February 19, 2025)

I. Dataset Description

This study utilizes the Medical Information Mart for Intensive Care (MIMIC-IV), a publicly available database containing de-identified health-related data from over 50,000 ICU admissions at the Beth Israel Deaconess Medical Center between 2008 and 2019. MIMIC-IV includes detailed information on demographics, vital signs, laboratory measurements, treatment interventions, and clinical outcomes. Its extensive coverage and diversity allow for developing predictive models and examining their performance across demographic groups, enabling analysis of potential biases and fairness. The dataset's granularity and diversity provide an opportunity to explore how AI models perform across demographic groups and assess disparities in predictive outcomes. [6] [7] [1]

Research Questions

This project aims to address the following **primary research question**:

How can AI models be developed and evaluated to ensure fairness in ICU mortality prediction concerning MIMIC-IV sensitive attributes?

The following **supporting sub-questions** will guide this research:

1. What sensitive attributes in the MIMIC-IV dataset contribute to biased outcomes in ICU mortality prediction?
2. How do Logistic Regression, XGBoost, and Multilayer Perceptron (MLP) models differ in predictive performance and fairness across demographic groups?
3. Which fairness metrics are most effective for evaluating biases in ICU mortality prediction models?
4. What bias mitigation strategies can improve model fairness without compromising predictive accuracy?

Objectives

This project aims to develop predictive models using Logistic Regression(LR), XGBoost, and Multilayer Perceptron (MLP) to forecast ICU mortality. It focuses on identifying potential biases within these models, specifically examining sensitive attributes such as race, gender, age, insurance type, ethnicity, language preference, marital status, and socioeconomic status. To evaluate model fairness, the project will utilize fairness metrics, including demographic parity, equal opportunity, equalized odds, disparate impact, and calibration across subgroups. Additionally, performance metrics like AUC(min), AUC(macro-avg), and AUC(minority) will assess disparities, with equalized odds difference measuring consistency in true and false positive rates.

The project also proposes bias mitigation strategies to improve model fairness while maintaining predictive accuracy. These strategies include reweighting (adjusting training sample weights), threshold modification (adjusting decision thresholds), and the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalances. Ultimately, this research seeks to advance the development of equitable AI systems in healthcare, ensuring predictive models support unbiased decision-making, optimize resource allocation, and reduce the cognitive burden on healthcare professionals. [5] [4] [2]

Data Availability

The data used in this study is available from the MIMIC-IV database (version 3.1), which can be accessed at https://physionet.org/content/mimiciv/3.1/. Access to the dataset requires credentialed approval through the PhysioNet platform after completing the required data use agreements and training on human subjects research.

II. Methodology

Data Preprocessing and Feature Engineering

The study employs a comprehensive data preprocessing pipeline implemented in Python using the Pandas library. The analysis incorporates six primary tables from the MIMIC-IV database, each contributing distinct clinical and demographic information:

1. Demographics of the patient (core/patients.csv)
2. Hospital admission records (core/admissions.csv)
3. Intensive Care Unit stay information (icu/icustays.csv)
4. Laboratory measurements (hosp/labevents.csv)
5. Vital signs and additional charted data (icu/chartevents.csv)
6. Patient output measurements (icu/outputevents.csv)

1. Feature Selection and Engineering

The study examines several sensitive attributes to assess potential algorithmic bias: - Demographic characteristics (age, gender) - Socioeconomic indicators (insurance type) - Cultural factors (language preference, ethnicity) - Social determinants (marital status)

All categorical variables undergo one-hot encoding to facilitate model compatibility and interpretation.

For clinical measurements, we implement a systematic aggregation approach:

1. Laboratory values: Statistical features including mean, standard deviation, minimum, and maximum values are computed to capture the distribution and range of measurements
2. Vital signs: Similar statistical aggregations (mean, standard deviation, minimum, maximum) are performed to represent physiological trends
3. Output events: Cumulative sum and mean values are calculated to quantify patient output patterns

2. Target Variable Construction

The target variable for mortality prediction is determined by analyzing the relationship between patient death time and ICU discharge, specifically identifying cases where death occurred during the ICU stay. The binary classification outcome serves as the dependent variable for our supervised learning models.

3. Data Integration and Preprocessing Pipeline

The main preprocessing pipeline includes several critical steps:

1. Feature integration: Merging of multiple data sources while maintaining data integrity
2. Missing value imputation: Implementation of appropriate strategies for handling incomplete data
3. Feature scaling: Standardization of numerical features to ensure comparable scales
4. Dataset partitioning: Stratified splitting of data into training and testing sets while preserving class distributions

4. Future Feature Engineering Considerations

The methodology allows for future refinement through several potential measures: 1) Incorporation of additional clinical parameters 2) Refinement of aggregation methodologies 3) Integration of temporal feature extraction 4) Development of feature interaction terms

A. Technical Implementation Note

The preprocessing pipeline is implemented using Python's scientific computing stack, primarily utilizing the Pandas library for data manipulation and scikit-learn for preprocessing operations. This ensures reproducibility and maintains computational efficiency when handling large-scale healthcare data.

III. Related Works

MIMIC-IV Dataset Usage

The MIMIC-IV dataset has demonstrated significant utility across various clinical research applications. Zhang et al. [13] developed a mortality predictive model for ICU patients with primary pulmonary hypertension (PPH), incorporating multiple physiological parameters including age, breathing rate, blood characteristics, glucose levels, and SAPS II severity scores. Their model demonstrated robust discriminative capability between high- and low-risk patients, with validated clinical utility.

In the domain of pharmacovigilance, Wei et al. [11] conducted research on medication-related information extraction, focusing on identifying drug names and their associated adverse effects from medical records. Vistisen et al. [10] utilized the dataset to investigate the prevalence and temporal distribution of extrasystoles in septic ICU patients, specifically examining their potential as predictors of fluid responsiveness. Additionally, Tasnim and Mamun's research [9] evaluated the impact of feature selection methodologies on ICU mortality prediction models, demonstrating that Principal Component Analysis (PCA) enhanced predictive accuracy across multiple machine learning classifiers including Logistic Regression, Decision Tree, K-Nearest Neighbors, and Support Vector Machine algorithms.

Related Methods to Research Questions??

Model Performance Comparison Studies

Wei et al. [11] implemented a comprehensive system comprising two main components: named entity recognition (NER) and relation classification (RC). Their methodology involved comparing deep learning-based approaches, particularly BI-LSTM-CRF, against traditional machine learning methods such as conditional random fields for NER and support vector machines for RC. Similarly, Tasnim and Mamun [9] conducted comparative analyses of feature selection techniques, specifically employing PCA across various classification models including Logistic Regression, Decision Tree, K-Nearest Neighbours, and Support Vector Machine.

Bias and Fairness in Clinical Prediction Models

Recent research has increasingly focused on addressing algorithmic bias in clinical prediction models. Yang et al. [12] introduced an adversarial training framework designed to mitigate biases while maintaining high clinical performance, demonstrating generalizability across various outcomes, models, and fairness definitions through validation in multiple hospital cohorts. Sufian et al. [8] addressed algorithmic bias in cardiovascular risk prediction models by integrating fairness-aware algorithms, SCIR models, and interpretability frameworks. Their research demonstrated that bias mitigation techniques successfully improved fairness metrics while maintaining strong predictive accuracy. Furthermore, Gu et al. [3] investigated demographic biases in COVID-19 mortality prediction models, evaluating transfer learning's effectiveness in improving model fairness across diverse racial and ethnic groups, with particular success in enhancing predictive performance for Non-Hispanic Black, Hispanic/Latino, and Asian populations.

Novelty Analysis

Building upon existing literature in clinical prediction models and fairness-aware machine learning, this research addresses critical gaps in the comprehensive analysis of fairness-performance trade-offs in ICU mortality prediction. While prior work by Yang et al. [12] and Gu et al. [3] explored bias mitigation through adversarial training and transfer learning approaches, these studies primarily focused on racial bias or theoretical frameworks, without utilizing the MIMIC dataset's rich clinical data.

This research advances the field through three key contributions. First, it provides a novel comparative analysis of fairness-performance trade-offs across Logistic Regression, XGBoost, and Multilayer Perceptron architectures in ICU mortality prediction. Second, it implements a comprehensive evaluation framework incorporating multiple fairness metrics, surpassing the typical one or two metrics used in existing studies. Third, it introduces an innovative end-to-end framework that integrates sensitive attribute analysis, model comparison, fairness evaluation, and bias mitigation strategies—components often addressed separately in current literature.

Through this integrated approach, the research not only extends existing work in fairness-aware clinical prediction models but also addresses fundamental gaps in methodology and practical implementation, contributing to the development of more equitable clinical decision support systems.

[1] Nikos Afxentis. Predicting mortality and algorithmic fairness of icu patients. Master's thesis, 2024.

[2] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215–e220, 2000.

[3] T Gu, W Pan, J Yu, et al. Mitigating bias in ai mortality predictions for minority populations: a transfer learning approach. *BMC Medical Informatics and Decision Making*, 25(1):30, 2025.

[4] A. E. W. Johnson, L. Bulgarelli, L. Shen, and et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[5] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv (version 3.1). PhysioNet, 2024.

[6] Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

[7] Anand Murugan. Implementing fairness in real-world healthcare machine learning through datasheet for database. Master's thesis, University of Waterloo, 2024.

[8] MA Sufian, L Alsadder, W Hamzi, S Zaman, ASMS Sagar, and B Hamzi. Mitigating algorithmic bias in ai-driven cardiovascular imaging for fairer diagnostics. *Diagnostics*, 14(23):2675, 2024.

[9] N Tasnim and SA Mamun. Comparative performance analysis of feature selection for mortality prediction in icu with explainable artificial intelligence. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE, 2023.

[10] ST Vistisen, M Xu-Wilson, C Potes, J Enevoldsen, and A Artigas. Prevalence and temporal distribution of extrasystoles in septic icu patients: The feasibility of predicting fluid responsiveness using extrasystoles. *Critical Care Research and Practice*, 2018:1–8, 2018.

[11] Q Wei, Z Ji, Z Li, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, 27(1):13–21, 2020.

[12] J Yang, AAS Soltan, DW Eyre, Y Yang, and DA Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digital Medicine*, 6(1):55, 2023.

[13] CY Zhang, YS Hu, ZY Meng, et al. Development and validation of a mortality predictive model for icu patients with primary pulmonary hypertension. *Scientific Reports*, 14(1):31497, 2024.