

Predicting Acute Kidney Injury in Septic Patients Using Logistic Regression with MIMIC-III Data

Antonio García Tierno,
Daniel Girbes Sardaña,
Ravneet-Rahul Sandhu Singh

Master in Health Data Science — MHEDAS

January 23, 2025



Table of Contents

1. Introduction	3
2. Methodology	4
Cohort Building	4
Column Mappings and eGFR Calculation	5
NaN Removal	6
Feature Selection	7
3. Results	9
Model Performance Metrics	9
Feature Contribution Analysis	10
Benchmarking Against Existing Studies	11
4. Conclusions	12

Introduction

- **Acute Kidney Injury (AKI):**

- Sudden decline in kidney function.
- Affects approximately 14% of hospitalized patients globally.
- Higher prevalence in ICU.

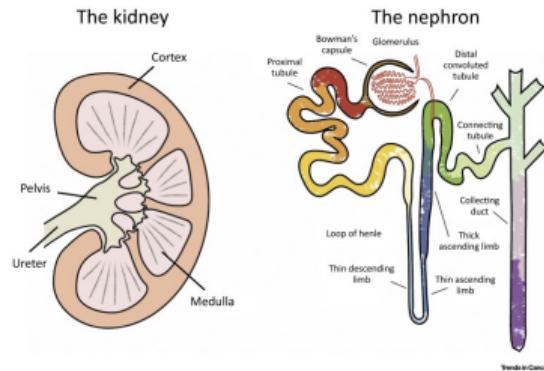


Figure: Kidney and nephron anatomy.



Figure: Kidney with AKI showing pale cortex and dark medullary tissue.

Cohort Building

Initial Cohort

ICD9_CODE IN ('99591', '99592', '78552')
Age: 18-89
LOS \geq 48 hours
Gender, Ethnicity, Age

Comorbidities

Acute Kidney Injury (AKI),
Chronic Kidney Disease (CKD),
Coronary Artery Disease
(CAD), Hypertension (HYP),
Type 2 Diabetes Mellitus (DM2)

Laboratory Features

Albumin, Anion Gap, Bilirubin, Blood Urea Nitrogen (BUN), Chloride, Creatinine, estimated Glomerular Filtration Rate (eGFR), Glucose, Hematocrit, Hemoglobin, International Normalized Ratio (INR), Lactate, Partial Thromboplastin Time (PTT), Platelet Count, Potassium, Sodium



Chartevent Features

Diastolic Blood Pressure (DBP), Heart Rate (HR), Height, Oxygen Saturation (SpO2), Systolic Blood Pressure (SBP), Temperature, Weight Mechanical Ventilation Vasopressor Use

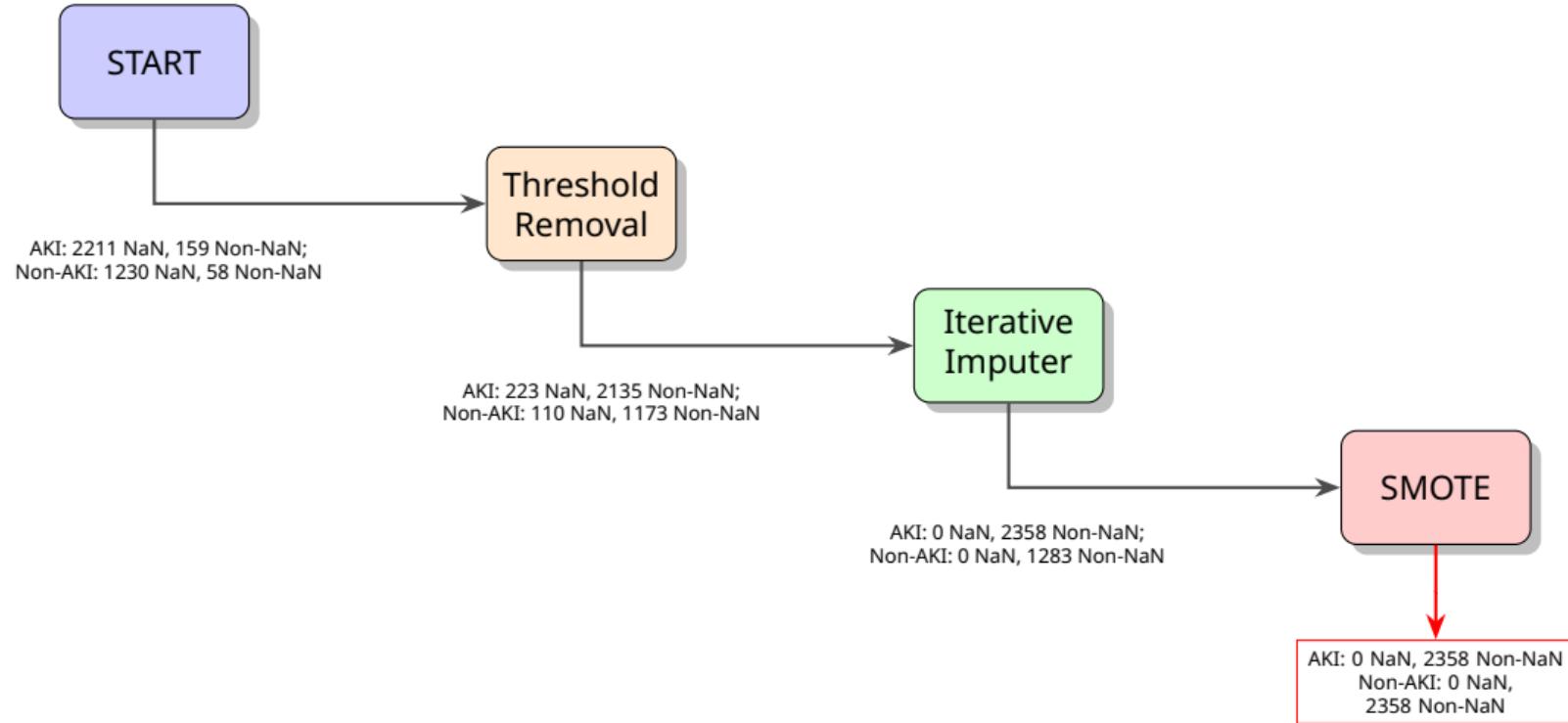
Column Mappings and eGFR Calculation

- Gender values = Male: **1**, Female: **0**.
- Ethnicities = **1**: Top1, **2**: Top2, **3**: Rest.
- All columns set to datatype **float** except for Gender, Ethnicities, and Comorbidities, which were **int**.
- eGFR calculated using a function provided in the **MIMIC-III GitHub** repository.
- Removed **Mechanical Ventilation** and **Vasopressor Use** columns as they were always 0 and provided no information.

```
1 def egfr(creat: float, age: float, gender: int, ethnicity: float) → float:  
2     """  
3         Calculate the estimated glomerular filtration rate (eGFR).  
4     """  
5     Inputs:  
6         - creat (float): Creatinine level.  
7         - age (float): Age in years.  
8         - gender (int): Gender.  
9         - ethnicity (float): Ethnicity.  
10    Outputs:  
11        - float or None: The calculated eGFR value or none.  
12    """  
13    # Return None if inputs are not correct  
14    if pd.isnull(creat) or creat == 0.0 or age == 0.0:  
15        return None  
16    # Calculate eGFR  
17    factor_gender = 0.742 if gender == 0 else 1  
18    factor_ethnicity = 1.212 if ethnicity == 2 else 1  
19    eGFR = 175 * (creat**-1.154) * (age**-0.203) * factor_gender * factor_ethnicity  
20    return eGFR
```

Figure: eGFR function.

NaN Removal



- Features were normalized to the range [0, 1] using the Max-Min function:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- Variance Inflation Factor (VIF) was calculated for all features using the formula:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

- Features with $\text{VIF} > 10$ were removed iteratively. The process continued until all features had $\text{VIF} < 10$.
- The t -test was used to compare the means of quantitative features between the AKI and Non-AKI cohorts.
- χ^2 -test to assess associations between categorical features and AKI status.

Feature Selection

Table: Comparison of features between AKI and Non-AKI patients. Quantitative features are summarized by mean and standard deviation (SD), while categorical features are presented as counts for each group. A feature is considered statistically significant if the *p*-value is ≤ 0.05 .

Feature (Unit)	AKI Mean (SD) / Count	Non-AKI Mean (SD) / Count	p-value
Age (years)	65.457 (14.986)	62.623 (15.045)	1.002×10^{-10}
Minimum Creatinine (mg/dL)	1.930 (1.513)	1.387 (1.753)	1.441×10^{-29}
Maximum Anion gap (mmol/L)	16.749 (4.605)	14.624 (3.604)	1.476×10^{-67}
Minimum Anion gap (mmol/L)	14.186 (3.753)	12.650 (3.070)	4.015×10^{-52}
Maximum Chloride (mmol/L)	108.420 (7.428)	107.488 (6.011)	2.234×10^{-6}
Minimum Chloride (mmol/L)	105.122 (7.557)	104.669 (5.856)	2.144×10^{-2}
Maximum Glucose (mg/dL)	174.921 (92.680)	155.957 (67.823)	1.341×10^{-15}
Maximum Platelet Count ($10^3/\mu\text{L}$)	225.550 (137.671)	237.127 (142.619)	4.588×10^{-3}
Maximum Potassium (mmol/L)	4.504 (0.822)	4.243 (0.683)	4.955×10^{-32}
Minimum Potassium (mmol/L)	3.862 (0.636)	3.707 (0.527)	1.777×10^{-19}
Maximum Sodium (mmol/L)	140.210 (6.018)	139.444 (4.198)	4.048×10^{-7}
Minimum Sodium (mmol/L)	137.386 (5.812)	137.058 (4.498)	3.030×10^{-2}
Minimum Hematocrit (%)	29.201 (5.483)	29.885 (4.839)	5.719×10^{-6}
Maximum Hemoglobin (g/dL)	10.647 (1.869)	10.867 (1.668)	2.086×10^{-5}
Minimum eGFR (mL/min/1.730 m ²)	54.405 (43.681)	97.813 (80.100)	5.736×10^{-112}
Minimum BUN (mg/dL)	39.019 (26.021)	20.502 (14.097)	2.487×10^{-185}
Minimum SBP (mmHg)	84.855 (14.873)	86.624 (13.538)	1.976×10^{-5}
Minimum DBP (mmHg)	41.174 (10.049)	42.601 (9.722)	7.452×10^{-7}
Maximum Temperature (°C)	37.611 (0.981)	37.858 (0.925)	7.728×10^{-19}
Minimum Temperature (°C)	36.077 (0.856)	36.221 (0.781)	1.383×10^{-9}
Gender	{0: 962, 1: 1396}	{0: 1172, 1: 1186}	9.689×10^{-10}
Ethnicity	{1: 1693, 2: 260, 3: 405}	{1: 1651, 2: 349, 3: 358}	2.707×10^{-4}
DM2	{0: 1382, 1: 976}	{0: 1597, 1: 761}	1.044×10^{-10}
CAD	{0: 1691, 1: 667}	{0: 1853, 1: 505}	5.794×10^{-8}
CKD	{0: 1593, 1: 765}	{0: 1972, 1: 386}	1.355×10^{-37}
HYP	{0: 1209, 1: 1149}	{0: 1451, 1: 907}	1.473×10^{-12}

Model Performance Metrics

Table: Performance metrics for training and testing datasets.

Metric	Train	Test
Accuracy	0.737	0.741
Precision	0.751	0.770
Recall	0.716	0.673
F1-score	0.733	0.718

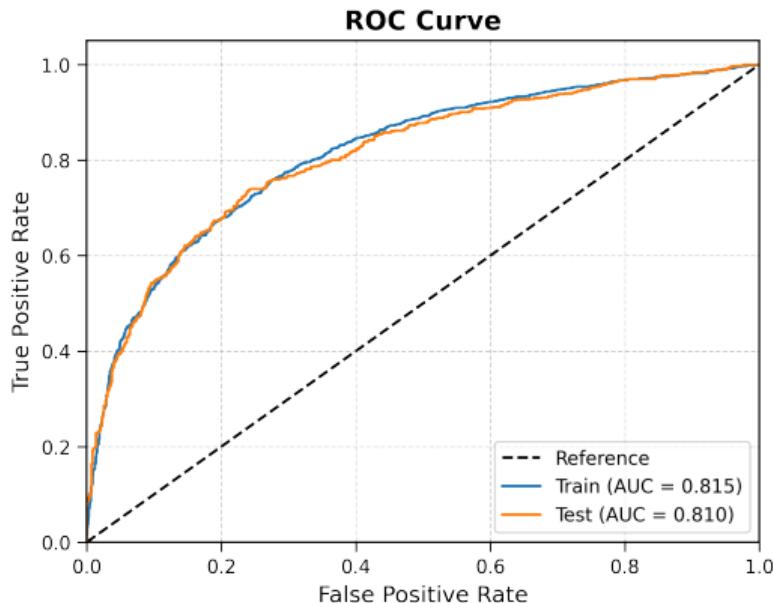


Figure: ROC curves for training and testing datasets.

Feature Contribution Analysis

- eGFR (most important marker for kidney function).
- Creatinine and BUN (indicators of renal health).
- Others: age, electrolytes, hemoglobin.

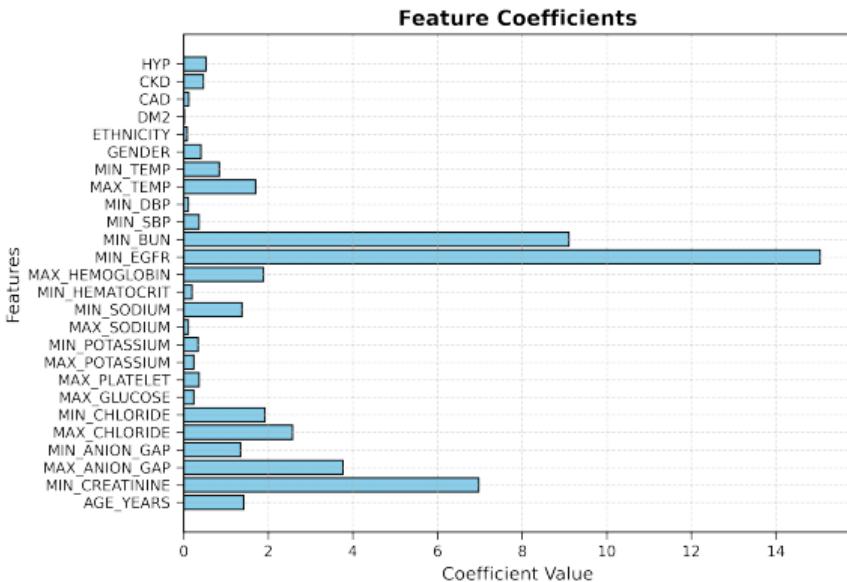


Figure: Feature importance in the logistic regression model.

Benchmarking Against Existing Studies

- Roknaldin et al. (2024): MIMIC-III database; 3301 ICU patients with sepsis; analyzed demographics, labs, vitals, and interventions. Multiple models.
- Malhotra et al. (2017): Multicenter dataset; 717 ICU patients; developed and validated an AKI risk score using logistic regression.
- Jiang et al. (2023): MIMIC-III database; 963 ICU patients with acute pancreatitis; developed a nomogram for AKI prediction using logistic regression.

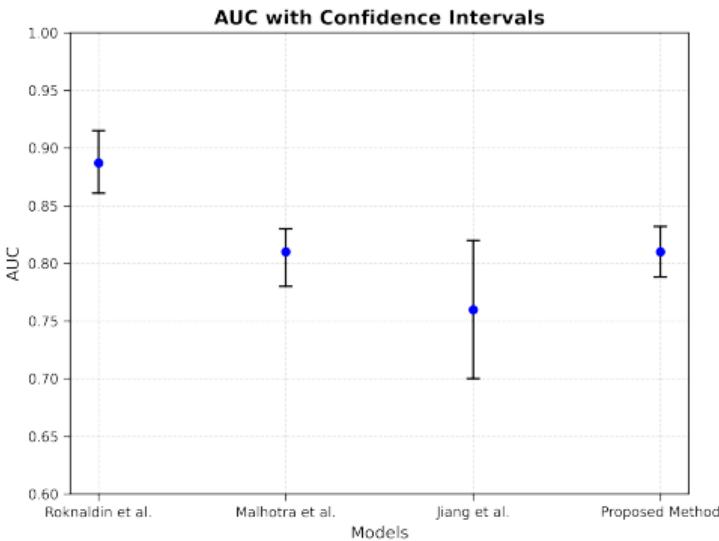


Figure: Comparison of AUC values across studies using logistic regression for AKI prediction.

- **Study Findings:**

- Developed a logistic regression model for AKI prediction.
- Achieved AUC = 0.810 (95% CI: 0.788–0.832), accuracy = 74.06%, precision = 76.97%, recall = 67.34%, and F1-score = 71.83%.
- Key predictors: eGFR, creatinine, BUN; secondary: age, electrolytes

- **Limitations:**

- Single-source data limits generalizability.
- Logistic regression may miss non-linear patterns.
- Missing data imputation introduces uncertainty.

- **Future Directions:**

- Add features like ICU interventions to improve accuracy.
- Validate on external datasets for broader applicability.
- Explore advanced models for capturing complex relationships.

Predicting Acute Kidney Injury in Septic Patients Using Logistic Regression with MIMIC-III Data

Antonio García Tierno,
Daniel Girbes Sardaña,
Ravneet-Rahul Sandhu Singh

Master in Health Data Science — MHEDAS

January 23, 2025

