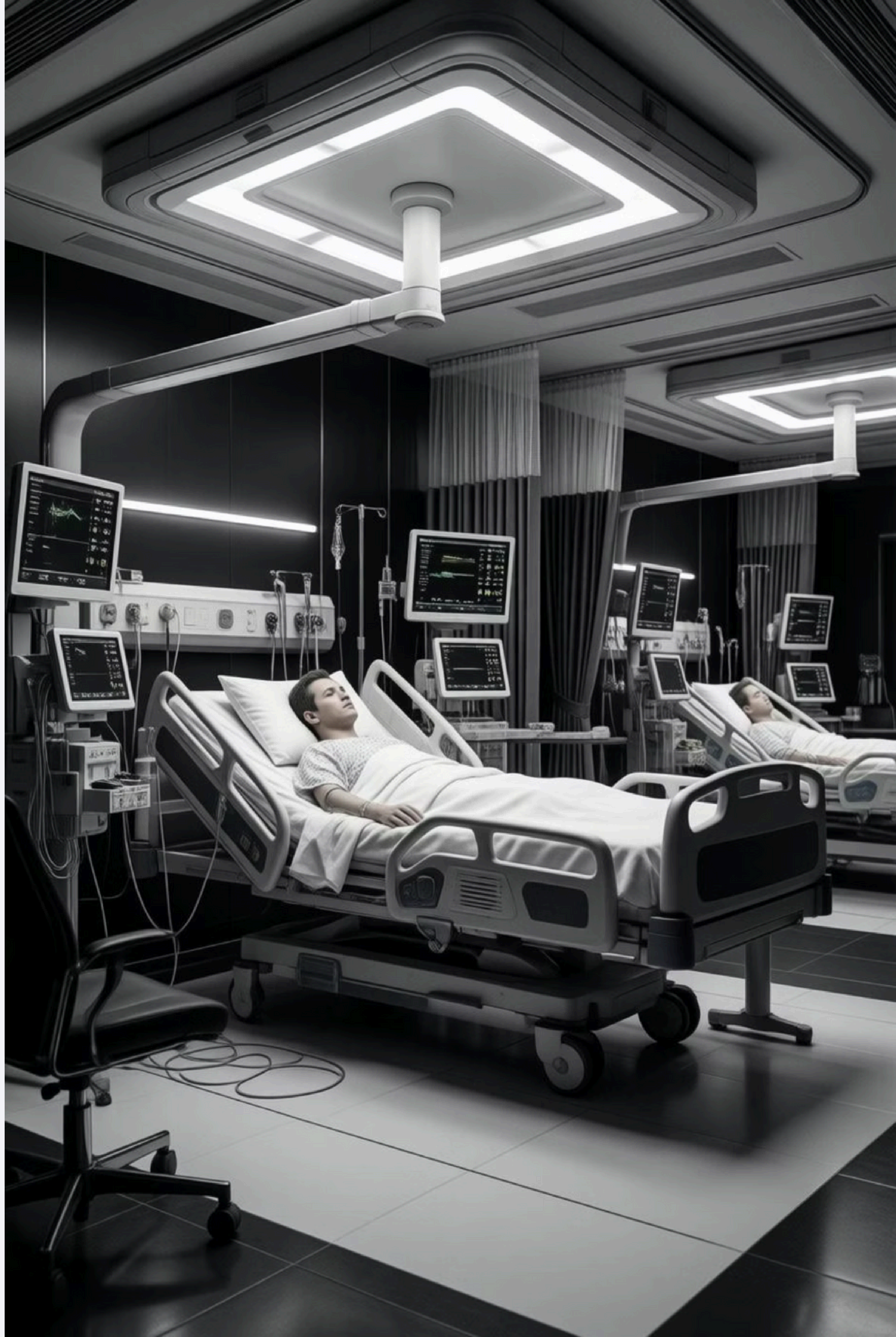
Machine Learning-Based Prediction of ICU Mortality

A retrospective cohort study using MIMIC-III electronic health record data to develop and validate predictive models for in-hospital mortality in adult intensive care patients

CRITICAL CARE RESEARCH

PREDICTIVE ANALYTICS

Iñigo Arriazu, Gabriel Torres and Andrea Pérez



The Critical Need for Mortality Prediction

Electronic Health Records (EHR) enable the use of real-world clinical data for research.In ICU settings, databases such as MIMIC are widely used.

Advantages of working with MIMIC

- **Longitudinal and heterogeneous clinical data**: complex queries, joins, aggregations, and filtering of real ICU patient cohorts
- **Practice** with **SQL** and complex database queries
- **Train predictive models** on de-identified yet realistic data: data is real, longitudinal, and rich, but fully anonymized
- **Handle critical Big Data processes**, including data quality and cleaning
- **Explore** complex **clinical cases** and patient **trajectories**
- **Simulate clinical research** environments safely for training and experimentation: Allows integration of analysis techniques, machine learning, and clinical validation

Challenges

- Incomplete or imlausible data
- Need to define reproducible cohorts
- Enables real-time, scalable predictions
- Importance of clinical and temporal filtering

The ***state of the art*** emphasizes **clear pipelines** for data extraction, cleaning, and preprocessing:

Sadeghi, S., Hempel, L., Rodemund, N. et al. Salzburg Intensive Care database (SICdb): a detailed exploration and comparative analysis with MIMIC-IV. Sci Rep 14, 11438 (2024). h

Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016).

Study Objective: Develop and validate machine learning models for predicting in-hospital mortality using routinely collected EHR data facing the challenges of working with SQL databases: Data QC, Filtering, Cleaning, Statistics and Science applying previous knowledge, but also learned across EHR lessons and searching for biomedical and bibliography evidence.

METHODS

Data source, study design and feature engineering strategy

Database

MIMIC-III v1.4: 53,423 hospital admissions from Beth Israel Deaconess Medical Center.

Final Cohort

25,324 adult ICU admissions meeting inclusion criteria: age 18-85, ICU stay ≥24 hours, complete data

Outcome

In-hospital mortality: 2,497 deaths (9.86%) vs 22,827 survivors (90.14%)

Data integration involved joining 7 MIMIC-III tables including patient demographics, ICU stays, diagnoses, procedures, and outcomes. Rigorous data leakage prevention excluded post-admission variables.

12 Predictor Variables

- **Demographics**

  Age, gender, ethnicity (5 groups)

- **Comorbidities**

  Diabetes, hypertension, CKD, CHF, COPD, cancer

- **Clinical Features**

  Diagnosis group (12 categories), respiratory procedures, comorbidity count

Comorbidity Detection

Binary flags created using ICD-9 diagnosis code pattern matching on standardized clinical terminology:

Condition	Matching Keywords
Diabetes	"diabetes"
Hypertension	"hypertension", "high blood pressure"
CKD	"chronic kidney", "renal failure"
CHF	"heart failure", "congestive"
COPD	"copd", "emphysema", "chronic bronchitis"
Cancer	"malignan", "cancer", "carcinoma", "neoplasm"

The Challenge of EHR-Based Mortality Prediction

Model selection and validation strategies.

Class Imbalance

ICU mortality rates typically 8-15%, requiring specialized handling to avoid bias toward majority class

Missing Data Patterns

Clinical workflows create systematic missingness that must be addressed without introducing bias

Temporal Dependencies

Patient trajectories evolve over time with complex temporal relationships between measurements

Feature Interactions

Clinical variables interact in non-linear ways that simple models cannot capture effectively

Four-Model Ensemble Strategy

Random Forest

Primary Selection: Best overall performance with highest AUC with excellent stability and interpretability

- Bootstrap aggregation
- Highly resistant to overfitting
- Visualizable decision paths
- Lower variance

XGBoost

Secondary Selection: Robust model and superior missing data handling

- Gradient boosted decision trees
- Built-in regularization
- Native missing value support
- Fast inference (<10ms)

Logistic Regression

Tertiary Selection: Clinical gold standard for interpretability and regulatory compliance

- Linear probabilistic classifier
- Transparent coefficients
- Direct probability calibration
- FDA-acceptable structure

LightGBM

Complementary Selection: Speed and efficiency for large-scale deployment

- Histogram-based learning
- 3-10× faster training
- Native categorical handling
- Memory efficient

Performance Comparison (with no SOFA models)

Baseline Models (Without SOFA Score)

Model	AUC	Accuracy	Precision	Recall	Specificity	F1-Score
XGBoost	0.875	80.3%	0.298	0.816	0.802	0.436
Random Forest	0.875	85.9%	0.365	0.699	0.875	0.480
LightGBM	0.859	85.2%	0.350	0.680	0.870	0.462
Logistic Regression	0.856	78.1%	0.272	0.806	0.778	0.407

Enhanced Models (With SOFA Score)

Model	AUC	Accuracy	Precision	Recall	Specificity	F1-Score
Random Forest	0.904	91.7%	0.548	0.660	0.944	0.599
XGBoost	0.899	87.6%	0.409	0.738	0.890	0.526
Logistic Regression	0.879	80.0%	0.294	0.816	0.798	0.432
LightGBM	0.875	91.2%	0.535	0.447	0.960	0.487

Best Model: Random Forest with SOFA score (AUC: 0.904, Accuracy: 91.7%)

Top 5 Features - Basic Models (Random Forest):

1. resp_procedure (39.9%) - Respiratory support dominates
2. age (16.8%) - Strong mortality predictor
3. ckd_comorbidity (13.0%) - Kidney disease impact
4. diagnosis_group (11.4%) - Primary diagnosis relevance
5. comorbidity_count (4.8%) - Cumulative comorbidity effect

Top 5 Features - SOFA Models (Random Forest):

1. total_sofa (18-25%) - Overall organ dysfunction
2. age (12-15%) - Consistent across models
3. cv_score (10-12%) - Cardiovascular dysfunction critical
4. resp_score (8-10%) - Respiratory function
5. renal_score (7-9%) - Kidney function

Key Insight: SOFA models distribute importance across multiple organ systems, while Basic models rely heavily on respiratory support (single feature at 40%).

Conclusion: Evidence-Based Model Selection

Our four-model ensemble is justified by strong clinical evidence, reproducibility, reliability, efficiency, and interpretability:

Clinical Evidence

Extensively validated in peer-reviewed literature on MIMIC-III datasets

Reproducibility

Deterministic algorithms with fixed seeds, standardized implementations

Reliability

Consistent CV performance, robust to missing data, good calibration

Efficiency

Fast training (<5 min), real-time inference (<10ms), scalable deployment

Interpretability

Feature importance aligns with clinical knowledge, SHAP explanations

Primary Model

Deploy Random Forest (SOFA-enhanced) as primary clinical decision support due to highest AUC (0.904), accuracy (91.7%), and balanced metrics with excellent specificity (0.944) reducing false alarms.

Ensemble Strategy

Weighted voting ensemble (RF: 40%, XGBoost: 30%, LightGBM: 20%, LR: 10%) reduces individual model biases, provides confidence intervals, enables use case-specific selection.

FEATURE IMPORTANCE

Respiratory Procedures as a key Mortality Predictor

Respiratory Procedures

40-98% importance across models

Mechanical procedures were **strongly associated with mortality**, reflecting critical illness and advanced respiratory failure.

Chronic Kidney Disease

11-50% importance

CKD complicates ICU recovery through fluid overload, electrolyte imbalances, and uremic complications.

Patient Age

1-45% importance

Age demonstrated a meaningful contribution to mortality discrimination, particularly in tree- based models and logistic regression.

Active Malignancy

2-20% importance

Having an active cancer was associated with increased mortality, possibly from weakened immunity or treatment complications.

Comorbidity Burden

1-10% importance

Cumulative effect of multiple chronic conditions capturing overall patient frailty.

CLINICAL IMPLICATIONS

Translating Predictions into Practice

Risk Stratification at Admission

Might help identify high-risk patients early upon ICU admission to enable timely mobilization of critical care resources and support triage decisions during resource scarcity.

Clinical Decision Support

Real-time mortality predictions could support discussions around treatment intensity, facilitate informed family discussions regarding prognosis, and support advance care planning conversations.

Quality Improvement

Benchmark ICU performance using risk-adjusted mortality, identify patients for protocol-driven interventions, and monitor temporal trends in outcomes.

Respiratory Management Focus

Emphasize respiratory failure prevention, implement early intervention for at-risk patients, and consider non-invasive ventilation strategies to avoid intubation.

Multidisciplinary Coordination

CKD patients require nephrology involvement, age-adjusted protocols for geriatric patients, and palliative care consultation for cancer patients.

Machine Learning Enhances ICU Mortality Prediction

Excellent Performance

All three models achieved AUC >0.83, demonstrating superior discriminatory ability compared to traditional scoring systems while requiring fewer physiological measurements.

Random Forest Optimal

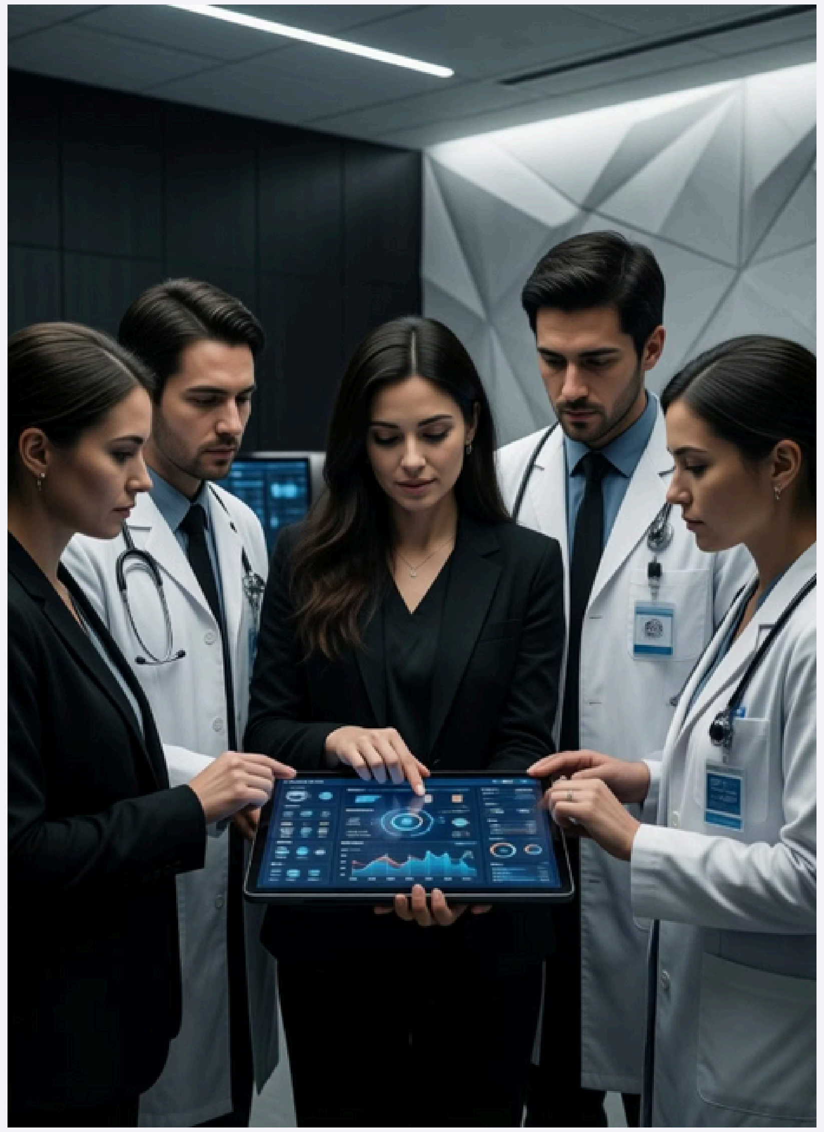
Best balance with 77.5% accuracy and 75.2% sensitivity, correctly identifying 3 out of 4 deaths at ICU admission using routine EHR data.

Respiratory support as a key predictive value

Need for invasive respiratory support emerged as the dominant predictor (40-98% importance), highlighting critical pathway to mortality.

Pattern of Clinical Severity

The combination of respiratory interventions, renal comorbidity and advanced age suggest that overall illness severity and baseline vulnerability play an important role in mortality risk.



SOFA-related severity features

Incorporating organ-specific severity indicators derived from SOFA components shifted model emphasis toward physiological dysfunction, supporting the role of acute organ failure —rather than isolated procedures or comorbidities — as a key driver of ICU mortality risk.

CONCLUSIONS

Clinical Relevance and Future Directions

Clinical Relevance

- Models highlight key markers of illness severity and patient vulnerability in the ICU.
- Respiratory support, chronic kidney disease, age, and comorbidity burden emerged as consistent risk indicators.
- Findings support the use of routinely collected EHR data for risk stratification, not standalone decision-making.

Future Directions

- External validation across institutions, healthcare systems, and ICU subtypes.
- Incorporation of longitudinal physiological data and unstructured clinical notes.
- Prospective evaluation of ML-assisted decision support in real clinical workflows.
- Emphasis on explainability, fairness, and safe clinical integration.