A Comprehensive Framework for ICU Outcome Prediction: Comparing Temporal Deep Learning and Gradient Boosted Models

Alon Bebchuk, Bar Vinizky-Shporn

August 14, 2025

1 Introduction and Motivation

We develop and evaluate a framework for predicting three critical ICU outcomes using the first 48 hours of hospital admission data: **(1) in-hospital mortality or death within 30 days post-discharge**, **(2) prolonged length of stay ($>$ 7 days)**, and **(3) 30-day hospital readmission post-discharge**.

Our approach uses GRU models with (X, M) tensor representation, where X contains feature values and M indicates missingness patterns. Key contributions include: systematic comparison of Multi-Task Learning (MTL) vs. Single-Task Models (STM), comprehensive multi-modal feature engineering, and thorough clinical evaluation including calibration, fairness, and interpretability analyses. We model outcomes as independent binary classification tasks, acknowledging the simplification of competing risks.

2 Data Extraction and Cohort Construction

2.1 Cohort Selection Criteria

The pipeline will begin with the list of `subject_id`s from the provided `initial_cohort.csv`. The final study cohort will be derived by applying the following inclusion/exclusion criteria:

1. **First Hospital Admission Only:** For each `subject_id`, only the chronologically first admission based on `admittime` will be considered. To establish a baseline model on a homogenous cohort and avoid confounding effects from heterogeneous prior admission trajectories, we restricted the study to each patient's first recorded hospital admission.

2. **Age at Admission:** Patients must be between 18 and 89 years old (inclusive). Ages for patients 90 and over are de-identified and shifted in the MIMIC-III database to protect anonymity; therefore, we cap the age at 89 to ensure data accuracy.

3. **Minimum Hospitalization Duration:** The length of stay must be at least 54 hours to ensure a valid 48-hour data window and a 6-hour prediction gap.

4. **Data Availability:** The admission must have associated chart events, as indicated by `has_chartevents_data = 1`.

5. **Valid Prediction Window:** Patients who died within the first 54 hours of admission will be excluded, as targets cannot be validly assessed at the 48-hour mark.

3 Feature Engineering Pipeline

Features described below will be extracted from data recorded strictly within the first 48 hours of admission.

3.1 Static Features (Admission-Level Context)

1. **Birth and Death Time:** The date of birth and death for the given patient. Helps to calculate the patient's age.

2. **Demographics:** Age, gender, ethnicity, insurance, marital status, language. Categorical features are one-hot encoded, with rare categories (<1%) collapsed into an "OTHER" bin.

3. **Admission and Discharge Time:** Provides the date and time the patient was admitted and discharged from the hospital.

4. **Admission Context and/or Location:** Admission type (e.g., 'EMERGENCY'), one-hot encoded, or information about the previous location of the patient prior to arriving at the hospital (e.g., 'EMERGENCY ROOM ADMIT').

5. **ICU Admission Time:** Compute the time until the first time a patient was transferred into the ICU. Include a binary flag if it happened in the first 48 hours.

6. **Patient Weight:** The first recorded admission weight (kg) within the 48-hour observation window is used. If no weight is recorded during this period, it is then imputed with the training set median.

7. **Major Interventions & Organ Support:** Static binary flags indicating receipt of critical interventions within 48 hours. The lists of `itemid`s and medication strings were compiled based on clinical expert review and established open-source pipelines (using `CPTEVENTS` and `PRESCRIPTIONS` tables). Features include:

   - `received_vasopressor`
   - `received_sedation`
   - `received_antibiotic`
   - `was_mechanically_ventilated`
   - `received_rrt` (Renal Replacement Therapy)

   Also includes binary flags for most microbiology events.

3.2 Time-Series Features (First 48 Hours)

1. **Vital Signs:** Heart Rate, Systolic/Diastolic BP, Respiratory Rate, SpO2, Temperature, and Glasgow Coma Scale (GCS) score. Multiple measurements within an hourly bin are aggregated using the mean.

2. **Laboratory Results:** Hemoglobin, White Blood Cell Count, Platelet Count, Sodium, Potassium, BUN, Creatinine, Glucose, Lactate, and INR.

4 Data Representation and Preprocessing

We use (X, M) tensor representation over 48 hourly bins, where $X_i = \{x_{i1}, \ldots, x_{i48}\}$ contains feature values and $M_i = \{m_{i1}, \ldots, m_{i48}\}$ indicates missingness ($m_{ij,d} = 1$ if missing, 0 if observed).

Processing: Hourly binning with mean aggregation, backward/forward-fill for missing values, global imputation with training mean for completely missing features, standardization, and final concatenation to shape `[batch_size, 48, D]`.

5 Model Architecture and Training

5.1 Models

MTL Model: Shared GRU backbone with three task-specific heads plus auxiliary reconstruction task for regularization. Uses uncertainty weighting:

$$\mathcal{L}_{\text{total}} = \sum_k \left(\frac{1}{2\sigma_k^2} \mathcal{L}_k + \frac{1}{2} \log(\sigma_k^2) \right) + \left(\frac{1}{2\sigma_{\text{recon}}^2} \mathcal{L}_{\text{recon}} + \frac{1}{2} \log(\sigma_{\text{recon}}^2) \right)$$

STM Models: Three independent GRU models, each with single task-specific head.

GBM Baseline: Three independent XGBoost models using flattened features (static + time-series summary statistics: mean, median, min, max, stddev).

5.2 Training

80/10/10 train/val/test split. Two-phase approach: hyperparameter search then evaluation with 3-5 random seeds.

6 Evaluation Protocol

Naming Convention: "FULL" indicates models use both feature values (X) and missingness patterns (M), otherwise only (X) is used. "RECON" indicates models include auxiliary next-timestep measurement reconstruction prediction.

Ablation Study: Compare 7 models: (1) GBM baseline, (2) STM-Baseline, (3) MTL-Baseline, (4) STM-Full, (5) MTL-Full, (6) STM-Full-Recon, (7) MTL-Full-Recon.

Metrics: AUROC/AUPR with 95% CIs, decision curve analysis, calibration (Brier score, reliability diagrams), SHAP interpretability, and fairness analysis across demographic subgroups.