Benchmarking Early Deterioration in Emergency Triage under Constrained Sensing *

KMA Solaiman Joshua Sebastian Karma Tobden
Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County
ksolaima@umbc.edu cj48611@umbc.edu ktobden1@umbc.edu

IEEE ICHI 2026

Motivation

A 7.0 magnitude earthquake strikes. Casualties floods in — no lab results, no history. *Who gets care first?*

Key problems in existing work:

- No labs, longitudinal data, or post-triage data in MCI
- Evaluation suffers from **patient-level leakage** and repeated-visit contamination
- No standardized benchmark contrasting hospital-rich vs. field-constrained (MCI-like) regimes
- **Lack** of public emergency triage datasets

Our Contribution:
A novel **patient-level, leakage-aware, reproducible benchmark** explicitly contrasting hospital-rich and vitals-only MCI-like settings using only first-hour data.

Dataset: MIMIC-IV-ED (10k Cohort)

Source: MIMIC-IV v3.1 + MIMIC-IV-ED v2.2 (PhysioNet, credentialed access)

Metric	Value	Notes
Unique patients	10,000	adults ($\geq$18y)
Primary outcome (y=1)	1,211 (12.1%)	in-hospital mortality
Secondary outcome	4,514 (45.1%)	ICU transfer $<$ 24 h
Age (mean $\pm$ SD)	62.5 $\pm$ 17.4	years
Male / Female	54.6% / 45.4%	

Primary outcome: In-hospital mortality; ICU transfer evaluated separately as a secondary endpoint.

Input window: Features restricted to the **first hour** of ED arrival — no longitudinal data, no post-triage interventions.

Same pipeline scales to the full MIMIC-IV-ED cohort ($\sim$425,000 encounters).

Two Triage Regimes

Two feature regimes reflect real-world information availability:

Feature Group	Hospital-Rich	MCI-Like
Vitals (T, HR, RR, SBP/DBP, SpO_2)	✓	✓
AVPU consciousness status	✓	✓
Derived resp./oxygen flags	✓	✓
Demographics (age, sex)	✓	—
Triage observations (pain, acuity)	✓	—
Chief complaint indicators	✓	—
Early labs (Hb, BUN, Na, K, Cr)	✓	—
Service utilization proxies	✓	—

- **Hospital-Rich:** Full first-hour picture — labs, vitals, observations, notes
- **MCI-like (Field):** Vitals, AVPU, and respiratory/oxygenation flags only — what a clinician has at point-of-first-contact

"How much do we lose when sensing is constrained to vitals only?"

Baseline Models

- **Logistic Regression (LR)** – Non-linear Interpretable baseline
- **Tree-based Ensembles:**
  - Random Forest (RF)
  - XGBoost (XGB)
  - LightGBM (LGBM)
- **TabNet** – attention-based; deep tabular neural network model

Leakage-Aware Benchmarking Framework

1. Load & Time-Align
First ED admission per patient; chunk-loading for scalability

↓

2. Initial Vitals
First complete vital set per (patient, stay); readmission-aware

↓

3. Admissions / ICU Linkage
Join via (subject_id, hadm_id); first ICU intime retained

↓

4. Labs
Earliest value within 1 h; pivot to wide format

↓

5. Observations & Proxies
AVPU from GCS verbal; resp. flags; FiO_2 estimation

↓

6. Label & Export
In-hospital mortality (or ICU $<$ 24 h); hospital-rich & MCI-like splits

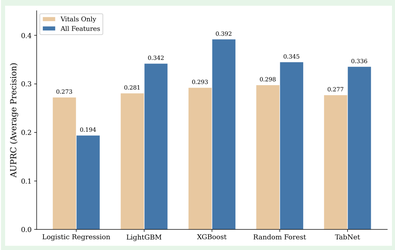
Leakage controls:

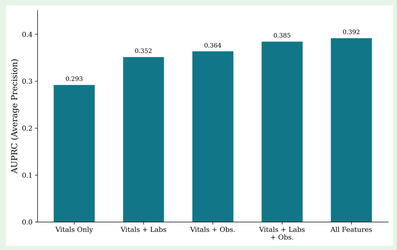
- **Patient-level splits** — no patient appears in both train and test
- **First-encounter only** — prevents repeated-visit contamination
- Imputation/scaling fit on training folds only and applied to test
- No post-triage features, downstream interventions, or outcome-adjacent variables

Results: Hospital-Rich vs. MCI-like

	Hospital-Rich		MCI-like (Vitals Only)	
Model	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
LR	0.668$\pm$.02	0.197$\pm$.01	0.704$\pm$.03	0.250$\pm$.04
LGBM	0.777$\pm$.01	0.358$\pm$.02	0.692$\pm$.02	0.263$\pm$.02
XGB	**0.809$\pm$.01**	**0.382$\pm$.03**	0.750$\pm$.01	**0.296$\pm$.02**
RF	0.771$\pm$.02	0.332$\pm$.04	**0.755$\pm$.01**	0.294$\pm$.02
TabNet	0.786$\pm$.03	0.334$\pm$.04	0.744$\pm$.01	0.271$\pm$.02

Average Precision under both regimes





- Performance gap is **consistent** but modest across all models
- Adding **observations** gives the largest single boost (+0.071 AUPRC)
- Labs add smaller but consistent gains; all-features reaches AUPRC = 0.392

Ablation: All Features vs. Vitals Only

AUROC (mean $\pm$ SD, 5 patient-level stratified splits):

Features	LR	LGBM	XGB	TabNet	RF
All Features	.668$\pm$.020	.777$\pm$.013	**.809$\pm$.011**	.786$\pm$.026	.771$\pm$.020
Vitals Only	.704$\pm$.032	.692$\pm$.015	.750$\pm$.011	.744$\pm$.010	**.755$\pm$.010**

- Performance gap is **modest** (ΔAUROC $\approx$ 0.06 for XGB)
- LR *improves* in vitals-only — richer features introduce collinearity it cannot exploit

Vital-Sign Ablation Study

Mortality as Deterioration, Leave-One-Out, AUROC

Features	XGB	TabNet	RF
All Features	**.809$\pm$.011**	.786$\pm$.026	.771$\pm$.020
Vitals Only	.750$\pm$.011	.744$\pm$.010	**.755$\pm$.010**
$-$RR	**.586$\pm$.013**	.510$\pm$.046	.570$\pm$.009
$-$SBP	**.588$\pm$.014**	.568$\pm$.036	.577$\pm$.014
$-$HR	**.599$\pm$.012**	.586$\pm$.038	.583$\pm$.005
$-SpO_2$	**.598$\pm$.016**	.575$\pm$.018	.579$\pm$.009
$-$Temp	**.591$\pm$.019**	.579$\pm$.012	.576$\pm$.016

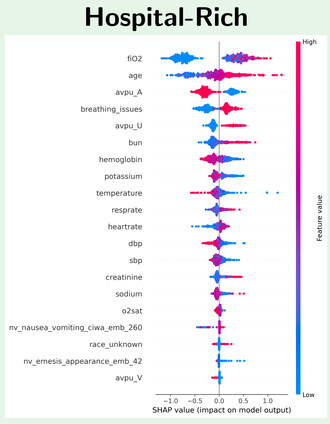
ICU Transfer Leave-One-Out (XGBoost, AUPRC)

Feature Set	AUPRC	ΔAUPRC
All features	0.815 $\pm$ 0.002	—
Vitals only	0.719 $\pm$ 0.001	0.000
Vitals $-$ SBP	0.678	$-$0.040
Vitals $-$ Temp	0.693	$-$0.026
Vitals $-$ RR	0.695	$-$0.024
Vitals $-$ HR	0.700	$-$0.019
Vitals $-SpO_2$	0.701	$-$0.018

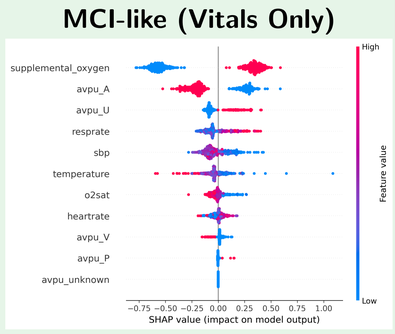
- **RR & Temp** removal causes largest drops in mortality prediction AUROC/AUPRC
- **SBP** is most critical for ICU transfer prediction
- HR contributes least in isolation across both outcomes

What Drives Predictions? (XGBoost)

Hospital-Rich



MCI-like (Vitals Only)



- **FiO_2, supplemental oxygen, AVPU, and RR** top both regimes
- Hospital-rich adds labs (BUN, Hb, K) as secondary signal
- MCI-like model reasoning is **fully preserved** — same signals, more compact
- SHAP rankings align broadly with ablation results

Key Insights & Future Work

- Vitals-only models retain **substantial discriminative power** — early physiological signal is strong
- Temperature and respiratory measures are the **most critical signals** for early triage risk
- XGBoost and RF are the strongest baselines; TabNet is competitive even without temporal data

Future Plans:

- Full-scale validation on $\sim$425k encounters
- Deep sequence baselines (RETAIN, AcuityNet) under partial-observability constraints
- MCI scenario simulation with structured sensor loss and deployed environments

Hospital-rich ED $\neq$ MCI-like Triage

• XGBoost achieves AUROC of 0.809 (hospital-rich) and 0.750 (MCI-like)
• Robust MCI-like Early Deterioration prediction
• Identifies necessary vitals for opportunistic-sensing in constrained triage environment

• Triage Benchmark • Deterioration prediction • MIMIC-IV-ED cohort • PhysioNet

*Accepted at the 14th IEEE International Conference on Healthcare Informatics (IEEE ICHI 2026)

H.A.R.M.O.N.I. Lab
Human-Aligned, Resilient, Multimodal, Open-world, Novelty-Informed Intelligence