Supplementary Materials: An Irregular Time-Aware Deep Learning Pipeline for Early Sepsis Prediction in Sparse ICU Trajectories

Annette Nayana Nellyet[1*], Manpreet Kaur[2] and Navdeep Kaur[3]

[1]University of Stirling, Ras Al Khaimah, UAE.
[2]Manav Rachna University, Faridabad, India.
[3]Mehr Chand Mahajan DAV College for Women, Chandigarh, India.

*Corresponding author(s). E-mail(s): annette.nellyet@gmail.com;
Contributing authors: manpreet@mru.edu.in; aulakh83@gmail.com;

Table S0: Cohort Attrition from MIMIC-IV Extract to Held-Out Test Set

The held-out test set referenced in Tables 1, 3, and 5 of the main text comprises 6,377 unique patients contributing 56,108 sliding windows (184 sepsis-positive patients). This count derives from a two-stage attrition: a patient-level filtering stage (Notebook 0, EDA and cohort construction) that reduces the raw 87,439-stay MIMIC-IV extract to the 76,307-stay analysis cohort, and a window-level filtering stage (Notebook 1, tensorization) that removes post-onset bins and windows falling inside the leakage exclusion zone $[t_0, t_0 + 6\text{h}]$ as defined in §3.1 of the main text. Table S0 walks through both stages.

Notes. The 15% allocation in the second stage is approximate because the patient-level random split partitions the 76,307 unique `subject_id`s with rounding; the exact post-split count is $\approx$ 11,446. The window-level filtering then removes (a) all bins occurring after t_{onset} within each ICU stay (so no model ever observes data recorded after onset), and (b) all sliding windows whose index time t_0 would generate a labelling window overlapping the immediate exclusion zone $[t_0, t_0 + 6\text{h}]$. Together these two window-level filters retain 6,377 of the $\approx$ 11,446 patients (a 56% retention rate; the other 44% contributed no valid 6-hour-ahead prediction windows under the leakage-prevention rules). Cohort-construction Table 1 in the main text reports per-group characteristics on the 76,307-stay filtered cohort: 67,727 non-sepsis controls and 8,580

Table 1 Table S0. Patient-level attrition from the raw 87,439-stay MIMIC-IV extract to the 6,377-patient held-out test set.

Stage	Stays	Change
Patient-level filtering		
Raw MIMIC-IV extract (adult ICU admissions)	87,439	—
of which left-censored sepsis (onset $\leq$ 6 h from ICU admission)	6,007	—
of which ICU-acquired sepsis (retained as positives)	11,978	—
After exclusion of left-censored sepsis cases	81,432	−6,007
After short-stay exclusion (< 24 h ICU length of stay)	76,307	−5,125
Train/val/test split + window-level filtering		
Patients allocated to held-out test set (15% patient-level random split)	$\approx$ 11,446	—
After post-onset bin removal and leakage-zone exclusion	6,377	—
Patients with ≥ 1 retained window in test set (final N)	**6,377**	—
of which sepsis-positive (contributed ≥ 1 positive window)	184	—
total retained test windows	56,108	—

sepsis stays (age 64.3 ± 17.0 vs. 66.2 ± 16.3 years, Mann–Whitney $p = 1.6 \times 10^{-21}$; 55.1% vs. 57.5% male, χ^2 $p = 3.7 \times 10^{-5}$; peak SOFA 3 [1–5] vs. 5 [3–7], $p < 10^{-100}$).

Table S1: Clinical Features and Bin-Level Missingness Statistics

We used 18 clinical variables in the raw feature space. The amount of missingness is shown for all time points in the filtered cohort (864,578 bin-level rows; 76,307 ICU stays). Features are depicted in the order that they are consumed by the network in the model input tensor.

Notes:

- Vital signs are recorded on a regular basis (missingness 1.8–28.3%), with temperature being the sparsest vital sign.
- The time series laboratory values are very sparsely observed (76.5–93.7% missing per 3-hour time bin), reflecting the bursty nature of laboratory draws in the ICU. Features such as bilirubin (93.7%) and lactate (88.6%) are the sparsest. With sparsity this extreme, we adopt time-aware features that model time since last measurement.
- Binary intervention indicators and demographics have zero missingness.
- The first-recorded SOFA score (`sofa_first`) — used for cohort grouping in analysis — is **not** included in the modeling features because it is the sole element that leaks the Sepsis-3 outcome definition based on the SOFA score.
- For the features, we use robust median/IQR scaling fitted only to the training set. Missing values are then set to the scaled training median, 0.0. This requires an observation mask M and an elapsed-time matrix Δt.

Table 2 Table S1. Feature list with bin-level missingness in the filtered MIMIC-IV cohort.

Idx	Feature	Category	Unit	Missing (%)
1	Heart rate	Vital sign	bpm	1.8
2	Mean arterial pressure (MAP)	Vital sign	mmHg	2.8
3	Respiratory rate	Vital sign	/min	2.7
4	SpO_2	Vital sign	%	2.6
5	Temperature	Vital sign	°C	28.3
6	Creatinine	Laboratory	mg/dL	77.4
7	Bilirubin	Laboratory	mg/dL	93.7
8	Platelets	Laboratory	$\times 10^3/\mu$L	78.4
9	White blood cell count	Laboratory	$\times 10^3/\mu$L	77.8
10	Sodium	Laboratory	mEq/L	76.9
11	Potassium	Laboratory	mEq/L	76.5
12	Lactate	Laboratory	mmol/L	88.6
13	Urine output	Derived clinical	mL/3h	25.8
14	GCS total	Derived clinical	3–15	31.1
15	Vasopressor active	Binary intervention	0/1	0.0
16	Mechanical ventilation active	Binary intervention	0/1	0.0
17	Age	Demographic	years	0.0
18	Gender (M=0, F=1)	Demographic	binary	0.0

Table S2: ATP Loss Hyperparameter Sensitivity Analysis

The two hyperparameters of the Asymmetric Temporal Penalty (ATP) loss are tuned in this experiment: the focal modulation exponent γ and the temporal urgency rate α. For this experiment a grid search was performed over $\gamma \in \{0, 0.5, 1.5, 3.0\}$ and $\alpha \in \{0, 0.1, 0.5, 1.0\}$. Due to time constraints a single seed (42) was used and only a reduced training budget was employed, using 10 epochs with a 3-step patience. The results of the grid search are shown in Table S2, and the settings used for the primary results that were run for all architectures are underlined. The test set was not used to select these hyperparameters. Note that the best validation loss was obtained with ($\gamma = 1.5, \alpha = 0.0$), with a validation loss of 0.623; however, this setting caused TD-GRU to be overly eager and achieve very low temporal coverage.

Notes:

- For PR-AUC maximum (validation maximum), we found ($\gamma = 1.5, \alpha = 1.0$) yields PR-AUC 0.0230. The relative gap on validation to the conservative ($\gamma = 1.5, \alpha = 0.5$) is 18.5%, or 0.0194 in value.
- *Justification for choosing $\alpha = 0.5$ as the primary configuration* — The validation-maximum advantage at $\alpha = 1.0$ did not translate to an advantage on the held-out test set, demonstrating only equivalent (to 3 d.p.) test PR-AUC of 0.0181 for $\alpha = 1.0$ versus 0.0184 for $\alpha = 0.5$ under full 5-seed training. However, this more aggressive configuration had much worse probability calibration (increased by 52% from 0.1034 to 0.1567 in terms of Brier score), and approximately doubled the seed-level standard

Table 3 Table S2. Validation PR-AUC across ATP loss hyperparameters. Bold indicates the validation maximum; underline indicates the conservative setting used for the primary analysis.

	Temporal decay α			
Focal γ	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1.0$
$\gamma = 0$	0.0179	0.0167	0.0189	0.0147
$\gamma = 0.5$	0.0166	0.0164	0.0167	0.0162
$\gamma = 1.5$	0.0164	0.0163	<u>0.0194</u>	**0.0230**
$\gamma = 3.0$	0.0162	0.0163	0.0194	0.0198

deviations for all metrics over the training set. The peak at $\alpha = 1.0$ in validation was viewed as slight overfitting from the single-seed hyperparameter search, and we consider the configuration ($\gamma = 1.5, \alpha = 0.5$) to be the appropriate primary configuration for ranking performance with equivalent lift but substantially better calibration and training-set stability.

- Both temporal-penalty columns ($\alpha = 0.5, 1.0$) substantially outperform the no-penalty columns ($\alpha = 0, 0.1$) when paired with focal modulation, confirming that the temporal urgency component improves ATP loss effectiveness.
- We observe that the interaction between temporal penalty (α) and focal down-weighting of easy negatives (γ) is non-trivial. Most surprisingly, for a strong temporal component ($\alpha = 1.0$) *without* focal down-weighting of easy negatives ($\gamma = 0$), we obtain the worst performance in the grid (0.0147). Thus, focal down-weighting of easy negatives is a necessary component for the temporal penalty to function correctly, especially under extreme class imbalance.
- The surface is flat across $\gamma \in \{1.5, 3.0\}$ and $\alpha \in \{0.5, 1.0\}$ with an average range of 0.0194–0.0230, meaning the variations of the performance are relatively moderate within this range of hyperparameters. Moreover, the rankings on the test set for $\alpha = 0.5$ and $\alpha = 1.0$ were found to be indistinguishable.
- These validation values come from training on a single seed (42) with a greatly reduced training budget of 10 epochs with patience 3. These values are intended to serve as order-of-magnitude indicators of relative sensitivity rather than final performance.

Table S3: Temporal Split — Detailed Per-Model Performance

For robustness to temporal distribution shift, models were re-trained on a within-MIMIC-IV temporal split by admissions year (all admissions $\leq$ 2169 for training, 2170–2177 for validation, $\geq$ 2178 for testing), yielding 53,645 training stays ($\approx$259,514 3-hour windows), 7,679 validation-set windows, and 73,246 test-set windows. Scaling was fit only on the temporal training split. All neural models were 5-seed and had the same training budget. Point estimates are taken directly from the temporal-split column of Table 6 in the main text. Bootstrap intervals were calculated from the averaged-prediction bootstrap with $n = 2{,}000$.

Table 4 Table S3a. Discrimination and calibration under temporal splitting (averaged 5-seed prediction vector; 95% patient-level cluster-bootstrap CI; $n = 2{,}000$). Brier point estimates are taken from the averaged-prediction vector.

Model	PR-AUC [95% CI]	ROC-AUC [95% CI]	Brier
TD-GRU (Proposed)	0.0140 [0.0116, 0.0177]	0.7310 [0.7060, 0.7549]	0.1090
GRU-D	**0.0150** [0.0121, 0.0204]	0.7213 [0.6961, 0.7464]	0.0940
Standard GRU	0.0142 [0.0118, 0.0179]	**0.7318** [0.7066, 0.7562]	0.1140
XGBoost	**0.0154** [0.0126, 0.0198]	0.7187 [0.6926, 0.7447]	**0.0091**

Table 5 Table S3b. Paired bootstrap comparisons on the temporal test set (TD-GRU vs. each baseline; $n = 2{,}000$). Δ PR-AUC is computed from the averaged-prediction point estimates in Table S3a.

Comparator	Δ PR-AUC	p (two-sided)
GRU-D	−0.0010	0.792
Standard GRU	−0.0002	0.916
XGBoost	−0.0014	0.541

No comparison reached significance ($p < 0.05$).

Table 6 Table S3c. PhysioNet 2019 utility score under temporal splitting.

Model	Best Threshold	Norm. Utility
XGBoost	0.1761	+**0.0367**
GRU-D	0.5200	+0.0238
TD-GRU (Proposed)	0.5124	+0.0224
Standard GRU	0.5737	+0.0098

Notes:

- All neural models perform roughly 15–24% worse than random splitting in terms of PR-AUC, which we attribute to temporal drift in documentation practices as well as shifts in case-mix across the years included in MIMIC-IV admission data.
- Under temporal splitting, GRU-D leads the neural architectures on PR-AUC (0.0150), while TD-GRU and Standard GRU follow at 0.0140 and 0.0142 respectively, with Standard GRU narrowly leading neural ROC-AUC (0.7318) over TD-GRU (0.7310). The rank order observed under random splitting is therefore not preserved for the lead neural architecture, but the differences between recurrent models on the temporal test set are not statistically significant (Table S3b), so within-group ranking should be interpreted as descriptive.
- XGBoost was the most robust model to temporal shift in this study, achieving a temporal-split PR-AUC of 0.0154, representing an increase of +0.0024 over its

random-split PR-AUC of 0.0130. It also achieved the highest utility under temporal splitting (+0.0367), suggesting that tree-based models may exploit temporal heterogeneity through implicit regularisation and warrant further investigation.

- Note that XGBoost has a much lower Brier score (0.0091) and that the recurrent model Brier scores (0.0914–0.1088) should be viewed in the context of the discrimination metrics and should be well above 0.54% prevalence. These uncalibrated recurrent model probabilities can however be post-hoc isotonic-regressed to recover the full predictive distribution. XGBoost performance on a measure such as Brier score should be viewed as indicating a well-calibrated output distribution at the prevalence-scale, and not as implying that it produces better probabilistic forecasts for all purposes.
- On the temporal test set, we did not find any comparisons to be significantly different (due to the small effective positive count and large distribution shift).