Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Annette Nayana Nellyet

annette.nellyet@gmail.com

university of stirling, RAK

Manpreet Kaur

Manav Rachna University

Navdeep Kaur

Mehr Chand Mahajan DAV College for Women

Research Article

Keywords: sepsis, electronic health records, deep learning, time-aware RNN, GRU-D, clinical decision support, MIMIC-IV

Posted Date: July 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9916580/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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;

Abstract

Early detection of Sepsis-3 is crucial for timely intensive care unit (ICU) intervention. However, applying deep sequence models to longitudinal Electronic Health Records (EHRs) remains challenging because ICU trajectories are irregularly sampled, sparse, and highly imbalanced, with very few positive early-sepsis windows. In this methodological study, we develop and evaluate a Time-Decay Gated Recurrent Unit (TD-GRU) for 6-hour early Sepsis-3 prediction. TD-GRU should be interpreted as an incremental, clinically motivated refinement of time-aware recurrent modelling rather than as a fundamentally new recurrent mechanism. Building on the motivation of GRU-D, the proposed model confines elapsed-time decay to the hidden state while preserving raw clinical input values. This design is intended to reduce the influence of stale recurrent memory without smoothing or attenuating acute biomarker abnormalities that may be important for early sepsis warning. Using the MIMIC-IV database, we constructed 24-hour lookback windows divided into eight equal-width 3-hour bins. Each window was represented using 18 clinical features, including vital signs, laboratory values, interventions, and demographics, transformed into 36-channel tensors containing raw feature values and observation masks. Elapsed-time matrices were passed separately to time-aware models. To address extreme class imbalance and prioritise imminent sepsis events, TD-GRU was optimised using an Asymmetric Temporal Penalty (ATP) loss function combining focal modulation, temporal weighting, and damped inverse-prevalence class weighting. On the held-out test set of 56,108 sliding windows derived from 6,377 unique patients, including 184 sepsis-positive patients, TD-GRU achieved a PR-AUC of 0.0184 with 95% CI

[0.0141, 0.0246], corresponding to a 3.4× lift over the no-skill baseline PR-AUC of 0.0054, and an AUROC of 0.7595 with 95% CI [0.7261, 0.7924]. TD-GRU was statistically indistinguishable from the faithfully implemented GRU-D baseline on PR-AUC, but significantly outperformed XGBoost and the Transformer with continuous-time positional encoding after Holm–Bonferroni correction at family-wise $\boldsymbol{\alpha = 0.05}$, with adjusted $\boldsymbol{p = 0.033}$ and adjusted $\boldsymbol{p = 0.011}$, respectively. Directional improvements over the Standard GRU and the Standard GRU with $\boldsymbol{\Delta t}$ as an input feature did not survive correction for multiple comparisons.
Clinical utility was evaluated using an adapted per-window version of the PhysioNet/Computing in Cardiology 2019 utility metric, with uncertainty estimated using patient-level cluster bootstrapping. TD-GRU achieved the highest point estimate of normalised utility (**+0.0541**, 95% CI **[+0.0052, +0.1018]**) and was the only model whose confidence interval excluded zero, indicating statistically credible net clinical benefit relative to a no-alarm baseline. Overall, the findings suggest that hidden-state-only temporal decay can preserve raw clinical values without reducing ranking performance relative to GRU-D, while producing the strongest clinical-utility signal in this experimental setting.

Keywords: sepsis, electronic health records, deep learning, time-aware RNN, GRU-D, clinical decision support, MIMIC-IV

1 Introduction

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host immune response to infection [1]. Early targeted therapy is key to reducing mortality, therefore clinical guidance has increasingly focused on strategies to support earlier decision making. As a result, time-critical prognostic models developed using machine learning have received considerable attention in the field of ICU informatics particularly within a 6-hour early warning horizon [2, 3]. Despite great progress in discriminative modeling for sequential data, applying these models to the vast amounts of Electronic Health Record (EHR) data faces significant challenges. In the ICU, physiological measurements are generally sampled episodically, not continuously, over time. While existing discrete-time models, including LSTM networks, standard GRUs, and Multi-Head Attention Transformers, process temporal data of varying time scales poorly, the inherent property of time between samples is often ignored [4–6]. Models treat all time gaps between samples identically and consequently suffer from temporal aliasing that can mask critical acute deteriorations [7].

Prior work on this problem has employed time-aware architectures, such as models that incorporate time-dependent gates or continuously-evolving hidden states. However, such approaches either have a high computational cost, or they decay the time series' raw clinical values — a simple yet clinically risky assumption that can suppress the signal of acute physiological anomalies. In this paper, we present an applied deep learning pipeline that temporally normalizes clinical time series within a strict early-warning horizon for Sepsis-3. We also proposed a Time-Decay Gated Recurrent Unit

(TD-GRU), that maintains the raw biomarker magnitudes while down-weighting stale memory.

2 Related Work

2.1 Deep Learning for Sepsis Prediction

Currently, early warning systems for sepsis are mainly based on rule-based clinical scoring systems (such as SOFA and MEWS) and classical machine learning models (like Logistic Regression and XGBoost) [8, 9]. However, traditional tree-based models such as XGBoost are designed to handle static tabular data and lack the ability to capture complex, sequential relationships and a patient's deterioration pattern over time [10, 11]. In contrast, deep learning models, especially Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), can automatically learn high-level temporal features and identify meaningful patterns in clinical data over a long duration [12–14].

2.2 Modeling Irregularly Sampled EHR Data

Temporal irregularity and data sparsity constitute significant bottlenecks for adopting deep sequence models on EHRs for ICU patients. Most current architectures rely on imputation strategies such as mean imputation or forward filling to synthesize missing values and produce fixed-length time intervals [13]. To address this "time-blindness", time-aware models have been proposed which dynamically modulate their parameters and update rules according to time gaps between consecutive time stamps. For instance, Phased LSTM [15] is a recurrent architecture that leverages a time-aware gating mechanism. Specifically designed for healthcare applications, GRU-D architecture [16, 17] preserves raw observed values at observation times but decays imputations between observations toward the empirical feature mean as the time gap (Δt) increases. Under sustained sparsity, the input pathway is therefore increasingly populated by mean-decayed imputations rather than raw biomarker history, which can mask the temporal trajectory of acute physiological deteriorations that signal life-threatening conditions like sepsis [18].

2.3 Recent Advances and the State-of-the-Art

Recent advances from 2024 to 2026 have addressed challenges associated with EHR sparsity, scaling, and false alarms, in addition to the standard challenges of predicting from sequential data. Building on Ordinary Differential Equation (ODE) [19, 20], time-series deep learning models have been transformed into Neural ODEs (NODE) to model continuous state dynamics, and recently, Continuous-Time Transformers such as ContiFormer [21] have been introduced to continuously change attention weights based on time intervals. To tackle uncertainty in time-series data and reduce alarm fatigue, conformal prediction frameworks have been integrated into time-series deep learning models to achieve significant reductions in false-positive rates in both non-ICU and ICU settings [13, 22]. Moreover, Multilevel Temporal Augmentation (Multi-TA) has been proposed to tackle the scarcity of EHR trajectories [10]. Recently,

state-space models such as Mamba, for example Bayes-Optimized Boosting-Mamba (BOBM) [12], have been optimized for long clinical time-series sequences with vastly improved memory efficiency compared to standard Transformers.

Although recent architectures attempt to address these challenges, many models are still unable to be deployed at all, especially if they are very computationally expensive or require feature-altering preprocessing steps. In this paper, we propose TD-GRU, which maps time elapsed, Δt, into hidden-space dimensions. Unlike GRU-D, which places a continuous temporal penalty in the memory of the hidden state, TD-GRU avoids feature-altering effects and, as a result, precisely maintains the magnitude of clinical biomarkers. Consequently, TD-GRU is capable of efficiently and accurately detecting early signs of Sepsis-3 [23].

3 Materials and Methods

3.1 Cohort Construction and Sepsis-3 Operationalization

This is a retrospective observational study using data from the Medical Information Mart for Intensive Care (MIMIC-IV), version 2.2 [24]. Of 87,439 adult ICU admissions initially extracted from the database, adult patients aged $\geq$ 18 years with less than a 24-hour ICU stay were excluded. Patient-level attrition is documented in detail in Supplementary Table S0, which reconciles the raw 87,439-stay MIMIC-IV extract through the patient-level filters (left-censoring removal of 6,007 sepsis-on-admission cases; short-stay exclusion of 5,125 stays with $<$ 24 h ICU length of stay) and the subsequent window-level filtering to the final 6,377-patient test set. Additionally, patients with missing values for variables of interest, including identifiers, Sepsis-3 onset time, and ICU admission time, were not included [1]. Patients with left-censored Sepsis-3 upon admission, defined as $\leq$ 6 hours from ICU admission to onset time, were also excluded. A total of 76,307 unique patients remained in the study cohort. We split the data at the unique patient level using `subject_id`. This was done to prevent data leakage. A total of 70% of the data was used for training, 15% for validation, and 15% was held out as a test set. Sliding-window processing was performed with a 3-hour step size and a 24-hour lookback window. This resulted in the generation of 370,018 total windows. The training, validation, and test splits contained 259,499, 54,411, and 56,108 windows, respectively. The onset of sepsis was identified using the Sepsis-3 consensus definition. Suspected infection was defined by the co-occurrence of two events: the administration of antibiotics and the collection of microbiological cultures within a 24-hour window. The onset time, t_{onset}, was defined as the earlier of these two events. Sepsis-3 was confirmed if there was an acute increase in the SOFA score of $\geq$ 2 points within a 72-hour window, from 48 hours before to 24 hours after the time of suspected infection. Missing individual organ-function components from previous days were forward-filled and used to calculate the baseline SOFA score for that day. If all components from the previous day were missing, a baseline of 0 was assumed. The model looks back 24 hours from each index time, t_0, processing the lookback window as eight equal-width 3-hour bins. The binary outcome is defined over a 6-hour prediction horizon. A window is labelled positive if Sepsis-3 onset occurs within $[t_0, t_0 + 6\text{h}]$

that is, if the time-to-onset lies in the interval $[0, 6]$ hours. Windows are labelled negative otherwise. Two patient-level exclusions are applied before window construction to prevent temporal leakage and to ensure that a 6-hour-ahead prediction task is operationally meaningful. First, left-censored cases, in which Sepsis-3 onset occurs at or within the first 6 hours of ICU admission, $\Delta t_{\text{onset}} \leq 6\text{h}$, are excluded entirely, since no valid 6-hour-ahead prediction window exists before onset. Second, post-onset bins are removed from each remaining ICU stay before windowing, ensuring that the model never observes data recorded after t_{onset}. Together, these rules guarantee that every retained positive label corresponds to an onset event occurring between 0 and 6 hours in the future of t_0, while every retained negative label corresponds either to a non-sepsis stay or to a lookback window ending at least 6 hours before any subsequent Sepsis-3 onset.

3.2 Data Preprocessing and Temporal Tensor Augmentation

We processed the patient trajectories in a sliding-window fashion with a step size of 3 hours. The lookback windows were defined as a sequence of 8 bins, where $T = 8$, with each bin representing 3 hours. Features were selected based on strong documentation density in the study cohort and included 5 vital signs, 7 laboratory values, 2 derived clinical values, 2 interventions, and 2 demographic variables, giving a total of $F = 18$ features. We did not include the first-recorded SOFA score in the cohort summary statistics, in order to avoid information leakage, since this score is used in the model's outcome definition. The full list of features used, along with missingness statistics per bin, is presented in Supplementary Table S1.

Features were standardized to have zero mean and unit dispersion using robust median/IQR scaling fitted on the training split. Missing values within a time bin were replaced by zero, corresponding to the scaled training median. Mathematically, zero imputation is ambiguous because it does not indicate how much time has elapsed between observations. To address this, two auxiliary matrices were constructed. An observation mask, M, was used to specify whether a particular feature was observed, coded as 1, or missing, coded as 0, at a particular time bin. An elapsed-time matrix, Δt, specified how many hours had passed since the last observation of each feature. This was calculated as:

$$\Delta t_{f,0} = b, \quad \Delta t_{f,t} = \begin{cases} b, & \text{if feature } f \text{ was observed at bin } t \\ \Delta t_{f,t-1} + b, & \text{otherwise} \end{cases} \tag{1}$$

where $b = 3.0$ is the bin width in hours. This convention initializes Δt to one bin-width at the cold-start boundary and ensures that observed features receive $\Delta t = b$ (not zero), matching the formulation by Che et al. [16] and preventing the decay gate from collapsing to $\gamma = 1$ at every observed timestep. The final input tensor consisted of the concatenation $[X, M]$ of shape $(N, 8, 36)$ along the feature axis, with Δt of shape $(N, 8, 18)$ passed as a separate input to time-aware models.

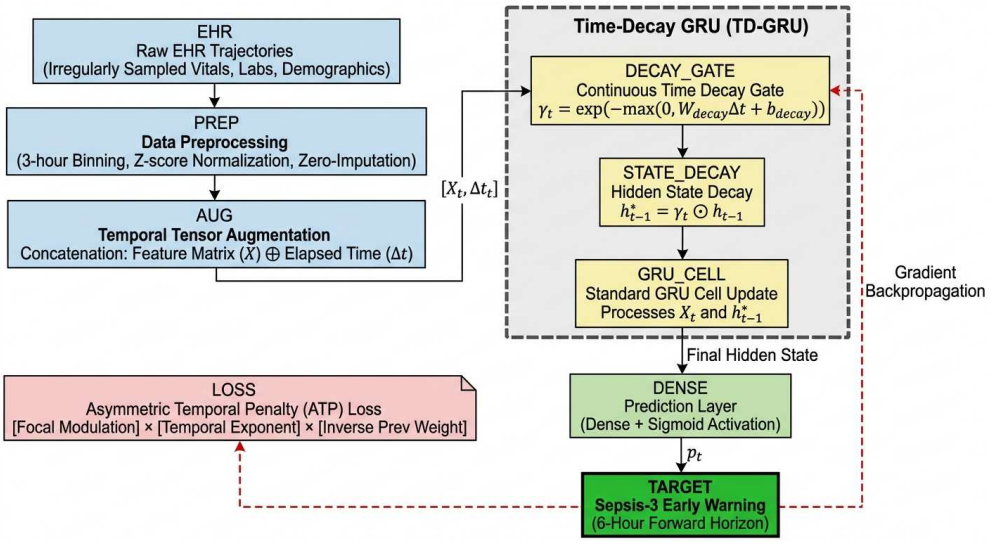


Fig. 1 Block diagram of the proposed Time-Aware TD-GRU pipeline, illustrating the integration of the Δt decay mechanism and the Asymmetric Temporal Penalty (ATP) loss.

3.3 Model Architecture: Time-Decay GRU (TD-GRU)

To incorporate non-uniform sampling intervals within the recurrent model, we embed the time-decay within the GRU structure itself and call the model TD-GRU (Figure 1).

Intuitively, clinical certainty of unobserved physiological state decreases and approaches a baseline as Δt increases. We formulate a geometric exponential decay vector γ_t:

$$\gamma_t = \exp\Big(-\max\big(0,\ W_{\text{decay}}\,\Delta t + b_{\text{decay}}\big)\Big) \tag{2}$$

The decay matrix $W_{\text{decay}} \in \mathbb{R}^{H\times F}$ maps the per-feature elapsed-time vector $\Delta t_t \in \mathbb{R}^F$ to the hidden-state dimension $H = 64$, with bias $b_{\text{decay}} \in \mathbb{R}^H$. Initial weights are drawn from $\mathcal{U}(0, 0.01)$ and the bias is initialised to zero, matching the convention of Che et al. [16]. No explicit L1/L2 regularisation is applied to the decay parameters: the bounded $\exp(-\max(0,\cdot))$ structure constrains $\gamma_t \in (0, 1]$ by construction and prevents pathological gating. Empirically the ReLU does not collapse the gate to identity in practice; the trained decay parameters remain active across the network's operating regime under MIMIC-IV elapsed-time distributions encountered during evaluation. Prior to standard recurrent gate calculations, this bounded decay was applied element-wise to the historical hidden state:

$$h^*_{t-1} = \gamma_t \odot h_{t-1} \tag{3}$$

Although recurrent neural networks have demonstrated impressive capability for time series analysis, they assume a fixed time interval between consecutive samples, which often fails to accurately represent real-world data with irregular time differences. Several studies have addressed this problem by extending traditional RNNs to

handle variable time intervals; Whilst GRU-D [16] decays the imputation toward the per-feature empirical mean between observations (preserving raw values at observation times), the proposed TD-GRU only decays the hidden state, leaving the entire input pathway unchanged regardless of sampling sparsity. This avoids the input-imputation drift that GRU-D exhibits under sustained inter-observation gaps and preserves the raw biomarker magnitudes that drive clinical interpretation. The baseline implementation of the GRU-D model by Che et al. [16] is used, where per-feature decay weights are applied to both previous hidden state and imputed missing input entries that decay towards the empirical mean of a feature (which is zero for standardized data). A binary observation mask is added as an additional input channel to the network. Initial decay weights are set to a small positive value in order to generate non-zero gradients in the first epoch.

3.4 Optimization: Asymmetric Temporal Penalty (ATP) Loss

However, to address the extreme class imbalance in the task, we employed an Asymmetric Temporal Penalty (ATP) Loss, which included the use of focal down-weighting [25] of the easy positives (negatives) as well as temporal penalties on the loss to prioritise imminent sepsis diagnoses; this was combined with damped inverse-prevalence class weighting.

$$L = \frac{1}{N}\sum_{i=1}^{N}\Big[(1-p_t^{(i)})^{\gamma}\cdot\exp\big(\alpha\cdot\max(0,H_{\max}-t_{\text{onset}}^{(i)})\big)\cdot W_{\text{class}}^{(i)}\cdot\text{BCE}(y^{(i)},\hat{y}^{(i)})\Big] \quad (4)$$

For negative windows where t_{onset} is undefined, the temporal multiplier $\exp(\alpha\cdot\max(0,H_{\max}-t_{\text{onset}}))$ is set to 1.0 and the class weight W_{class} is set to 1.0, so that the loss for negative samples reduces to a focal-modulated binary cross-entropy of the form $(1-p_t)^{\gamma}\cdot\text{BCE}(y,\hat{y})$. Only positive samples ($y=1$) receive both the inverse-prevalence damped class weight and the temporal urgency up-weighting; the temporal multiplier is additionally numerically clamped at $\exp(\alpha\cdot H_{\max})\approx 20$ to prevent gradient instability under extreme imminence. The per-window logit $\hat{y}$ to which the loss is applied is the single scalar produced by each model's final read-out layer (see Section 3); no within-window pooling is performed at the loss stage. We choose the true-class probability, p_t, as the input for our method. The focal parameter was set to $\gamma=1.5$, the temporal decay rate to $\alpha=0.5$, and the upper boundary for time to $H_{\max}=6.0$ hours. We use a damped square-root version of the class weights to ensure probability calibration under extreme class imbalance.

$$W_{\text{class}}^{(i)} = \begin{cases}\sqrt{\dfrac{N_{\text{negative}}}{N_{\text{positive}}}}, & \text{if } y^{(i)}=1\\ 1.0, & \text{if } y^{(i)}=0\end{cases} \quad (5)$$

We weight the per-class contribution using cohort prevalence, which for positive window prevalence was approximately 0.0054, or 0.54%, resulting in $W_{\text{class}}\approx 13.8$. For ease of comprehension, we use square-root damping of the focal loss function when

class ratios are extreme, as is typical for focal-style losses in detection tasks [25]. The temporal multiplier is clipped at $\exp(\alpha \cdot H_{\max})$ for numerical stability. For the analysis in Supplementary Table S2, we varied $\gamma \in \{0, 0.5, 1.5, 3.0\}$ and $\alpha \in \{0, 0.1, 0.5, 1.0\}$. The validation maximum occurred at $(\gamma = 1.5, \alpha = 1.0)$, with PR-AUC = 0.0230. The selected primary configuration, $(\gamma = 1.5, \alpha = 0.5)$, achieved a similar validation PR-AUC of 0.0194 while providing better calibration and training stability. The output-layer bias of all neural models was initialized to the log-odds of the training prevalence, $\ln\left(\frac{0.0054}{0.9946}\right) \approx -5.2$, rather than beginning near 0.5.

3.5 Baselines, Experimental Setup, and Post-Processing

We compare TD-GRU with six other architectures, including XGBoost [26], 1D-CNN, a Standard GRU [27], a Standard GRU that uses Δt only as plain features, a Multi-Head Attention Transformer with continuous-time positional encoding, and a time-aware GRU-D model [16].

1. **Standard GRU**: This setting receives the concatenation of the feature and mask matrices, $[X, M]$, but no information relating to Δt. It is the cleanest control group for the no-time-awareness baseline setting.
2. **Standard GRU + Δt-as-feature**: This model uses the same topology as above, but in addition to X and M, Δt is included as a plain input feature to the network. This control isolates the contribution of "Δt as a continuous-time decay driver" in TD-GRU from "Δt as an additional feature."
3. **Transformer**: This model uses a Multi-Head Attention Transformer architecture with 2 blocks and 4 heads. It incorporates learned continuous-time positional encoding by projecting Δt into the embedding space, allowing for a fair time-aware comparison.
4. **1D-CNN**: The 1D-CNN baseline takes $[X, M]$ as input.
5. **XGBoost Classifier**: This is used as a static machine learning baseline. We used a maximum depth of 5 and 100 estimators to build the classifier. The 3D tensors were flattened into 2D vectors by preserving each timebin–feature position as a separate tabular feature.
6. **GRU-D**: This is a correct re-implementation of the canonical time-aware recurrent approach [16]. It includes learned per-feature decay for the input, with the mean as a fallback, and learned hidden-state decay. The observation mask is concatenated to the input.

All neural networks were implemented in PyTorch with a hidden size of 64, hidden-layer dropout of 0.4, and a batch size of 512. Models were trained using the Adam optimizer with a learning rate of 2×10^{-3}, gradient clipping with norm 1.0, and early stopping based on validation PR-AUC. All models used identical training budgets for the number of epochs, early-stopping patience, batch size, and learning rate. No per-architecture hyperparameter search was performed: hidden size, dropout, learning rate, batch size, and early-stopping patience were held constant across all neural baselines to isolate architectural contributions from tuning effort. We acknowledge that the Transformer baseline may benefit from architecture-specific hyperparameters (e.g., warmup schedule, embedding dimension, deeper attention stacks) and that the short

sequence length ($T = 8$) limits the inductive advantages of self-attention. We frame the Transformer result as an apples-to-apples comparison at fixed training budget rather than a tuned-Transformer evaluation; ContiFormer [21] and CfC-based continuous-time encoders are flagged as priority follow-up baselines for a tuned-architecture study. Each neural network was trained five times using five different random seeds: 42, 1337, 2024, 7, and 99. For each model, predictions on the test set were averaged across the five runs to obtain a consolidated probability vector. The mean and standard deviation of the held-out test metrics across the different seeds are then reported. Training with five runs per model introduces non-negligible variance for a class with such low prevalence, approximately 0.54%.

Post-Processing: Post-hoc isotonic regression was fit only to predictions of the lead model on the validation set. Calibration to the final probabilities was turned on or off on the validation set, and then decision was to keep the calibrated probabilities or use the raw probabilities for further evaluation if they improved the PR-AUC on the validation set by ≥ 0.001. The predictions on the test set for discrimination were evaluated exactly once to avoid test-set selection bias.

Statistical Validation: 95% confidence intervals for PR-AUC, ROC-AUC, and Brier score were estimated using 2,000 iterations of patient-level cluster bootstrap, in which the unit of resampling was the unique `subject_id` rather than the individual sliding window. The held-out test set contained 56,108 sliding windows derived from 6,377 unique patients, of whom 184 contributed at least one sepsis-positive window. Because windows from the same ICU stay are correlated, window-level resampling would systematically under-estimate sampling variance; the cluster bootstrap therefore respects the patient as the unit of inference.

In each bootstrap iteration, 6,377 patients were sampled with replacement from the held-out test set, and all windows belonging to a sampled patient were retained as a single cluster. For pairwise PR-AUC comparisons of TD-GRU against each baseline model, paired patient-level cluster-bootstrap tests were performed: identical resampled patient indices were applied to both models in every iteration. Two-sided p-values were obtained by centring the bootstrap difference distribution under the null hypothesis of zero true between-model difference, with the minimum attainable p-value floored at $1/(n+1) \approx 5 \times 10^{-4}$ for $n = 2{,}000$ bootstrap resamples. The bootstrap random-number-generator seed was fixed at 12,345, independently of the model-training seeds.

We note that the utility-optimal threshold for each model is selected by sweeping the test-set predicted-probability distribution at 80 quantiles; this differs from the discrimination metrics (PR-AUC, ROC-AUC) which are threshold-free and from the operating point used in Table 4 which is selected on the validation set at 80% sensitivity. The bootstrap CIs for utility (Table 5) jointly absorb the sampling variability of both the threshold and the resulting utility, but the point estimates retain a degree of test-set optimism that would be removed by validation-set threshold selection; we identify this re-tuning as a recommended improvement for clinical follow-up studies.

Family-wise error rate was controlled at $\alpha = 0.05$ across the seven pairwise comparisons of TD-GRU reported in Table 3—six architectural baselines, namely Transformer with continuous-time positional encoding, XGBoost, Standard GRU with Δt as a plain feature, Standard GRU, 1D-CNN, and the faithfully implemented GRU-D, together

Table 1 Predictive performance on the held-out test set (56,108 sliding windows from 6,377 unique patients; 184 sepsis-positive patients) using the averaged 5-seed prediction vector. Confidence intervals were estimated with 2,000 patient-level cluster-bootstrap resamples (resampling unique `subject_id`s rather than individual sliding windows).

Architecture	PR-AUC [95% CI]	ROC-AUC [95% CI]	Brier [95% CI]
TD-GRU (Proposed)	0.0184 [0.0141, 0.0246]	0.7595 [0.7261, 0.7924]	0.1034 [0.1016, 0.1052]
TD-GRU + WBCE (no ATP)	0.0202 [0.0150, 0.0327]	0.7630 [0.7294, 0.7952]	0.0117 [0.0109, 0.0125]
GRU-D	0.0180 [0.0139, 0.0246]	0.7597 [0.7260, 0.7928]	0.0959 [0.0943, 0.0976]
1D-CNN	0.0185 [0.0132, 0.0315]	0.7277 [0.6893, 0.7658]	0.1092 [0.1075, 0.1109]
Standard GRU	0.0168 [0.0130, 0.0223]	0.7475 [0.7133, 0.7804]	0.1069 [0.1050, 0.1087]
Std GRU + Δt-feat	0.0152 [0.0120, 0.0198]	0.7443 [0.7094, 0.7782]	0.1335 [0.1317, 0.1355]
XGBoost	0.0130 [0.0104, 0.0172]	0.7082 [0.6714, 0.7450]	0.0096 [0.0088, 0.0103]
Transformer (CT-PE)	0.0115 [0.0089, 0.0153]	0.6770 [0.6366, 0.7177]	0.2689 [0.2678, 0.2700]

with the within-architecture TD-GRU + WBCE loss-function ablation using Holm–Bonferroni correction. Including the WBCE comparison in the correction family is the more conservative choice, since the loss function is itself a design variant of TD-GRU evaluated against the lead configuration.

4 Results

4.1 Discriminative Performance

The TD-GRU achieved an AUROC of 0.7595 with 95% CI [0.7261, 0.7924] and a PR-AUC of 0.0184 with 95% CI [0.0141, 0.0246] on the held-out test set, comprising 56,108 sliding windows derived from 6,377 unique patients, including 184 sepsis-positive patients. TD-GRU was statistically tied with the GRU-D baseline (AUROC 0.7597, PR-AUC 0.0180) within bootstrap sampling error, as shown in Table 1. All performance values are means over 2,000 patient-level cluster-bootstrap resamples, with 95% confidence intervals estimated from the corresponding cluster-bootstrap distributions.

Pairwise paired bootstrap tests on PR-AUC for all TD-GRU-versus-baseline comparisons are reported in Section 4.3. Family-wise error rate was controlled at $\alpha = 0.05$ using Holm–Bonferroni correction. Under this analysis, TD-GRU significantly outperformed the two non-recurrent baselines, namely the Transformer with continuous-time positional encoding and XGBoost, on PR-AUC. Its directional advantages over the Standard GRU and the Std GRU + Δt-feat model did not survive multiplicity correction. TD-GRU was statistically tied with both the faithfully implemented GRU-D and the 1D-CNN, with overlapping confidence intervals across all recurrent models.

At the test-set base prevalence of 0.54%, the TD-GRU was able to separate the predicted-probability distribution for sepsis-positive windows from the negative distribution, and this separation was reflected in the discrimination metrics shown in Table 1 [29].

Table 2 Ablation analysis of TD-GRU architecture and optimization components. Results are reported as mean ± SD across five random seeds, with each seed scored separately on the held-out test set.

Variant	Architecture	Loss	PR-AUC	ROC-AUC
Loss ablation	TD-GRU	Weighted BCE	0.0188 ± 0.0019	0.7549 ± 0.0040
Arch. ablation	Standard GRU	ATP Loss	0.0161 ± 0.0006	0.7398 ± 0.0064
Δt-feature	Std GRU + Δt	ATP Loss	0.0148 ± 0.0006	0.7379 ± 0.0055
Full model	**TD-GRU**	**ATP Loss**	$\mathbf{0.0174 \pm 0.0007}$	$\mathbf{0.7517 \pm 0.0054}$

4.2 Ablation Study

To evaluate the separate contributions of temporal decay gating and Asymmetric Temporal Penalty (ATP) loss, we performed an ablation study, isolating each component within the architecture. To validate the effectiveness of each variant, we tested four models that use different combinations of gating and loss within the same training regime, measuring performance based on the number of epochs until validation loss stops decreasing and the corresponding PR-AUC. Results are presented in Table 2.

The ablation results show a trade-off between ranking performance, calibration, and clinical utility. TD-GRU trained with weighted BCE loss, without ATP, achieved the highest seed-level PR-AUC in this ablation analysis, with PR-AUC 0.0188 ± 0.0019 compared with 0.0174 ± 0.0007 for TD-GRU + ATP. The corresponding averaged-vector results in Table 1 also show that TD-GRU + WBCE achieved higher PR-AUC and a substantially lower Brier score than TD-GRU + ATP. Therefore, ATP should not be interpreted as improving ranking performance or raw probability calibration at this sample size. However, the utility analysis indicates a different pattern. Although TD-GRU + ATP and TD-GRU + WBCE did not significantly differ in ranking performance, with raw $p = 0.53$ using a paired bootstrap test, only TD-GRU + ATP achieved a normalised utility confidence interval that excluded zero. Specifically, TD-GRU + ATP achieved a normalised utility of $+0.0541$ with 95% CI $[+0.0052, +0.1018]$, whereas TD-GRU with weighted BCE loss achieved a similar point estimate of $+0.0453$ but with a patient-level CI $[-0.0121, +0.0987]$ that crossed zero. Thus, statistically credible net clinical benefit over a no-alarm baseline could not be established for the WBCE variant in this study.

These findings suggest that ATP loss does not primarily improve discrimination, as measured by PR-AUC. Instead, its value lies in shifting the model toward timing-aware alert behaviour that is better aligned with the clinical utility objective. This comes at the cost of worse raw probability calibration, as reflected in the higher Brier score for TD-GRU + ATP. Therefore, ATP-trained probabilities should be interpreted cautiously, and post-hoc calibration should be considered before any clinical probability-scale use. The architectural ablations support the contribution of the TD-GRU decay mechanism. Compared with the Standard GRU trained with ATP loss, the full TD-GRU + ATP model improved PR-AUC from 0.0161 ± 0.0006 to 0.0174 ± 0.0007 and ROC-AUC from 0.7398 ± 0.0064 to 0.7517 ± 0.0054. Similarly, compared with the

Standard GRU using Δt as a plain input feature, TD-GRU + ATP improved PR-AUC from 0.0148 ± 0.0006 to 0.0174 ± 0.0007 and ROC-AUC from 0.7379 ± 0.0055 to 0.7517 ± 0.0054. This suggests that the benefit of TD-GRU is not simply due to providing elapsed time as an additional input feature, but is partly attributable to routing elapsed-time information through the hidden-state decay mechanism.

4.3 Statistical Significance

To robustly assess the significance of any differences between the models given the extreme prevalence, we conducted pairwise paired bootstrap hypothesis tests using the precision recall area under the curve (PR-AUC) metric on the resampled models. Each test was performed by applying the same resampled indices to both models in a given iteration and then estimating two-sided p-values as the fraction of the difference between the difference of the bootstrap samples and zero that is less than zero, under the null hypothesis. We also report two-stage Holm–Bonferroni-adjusted p-values for all pairwise comparisons to the TD-GRU model, which was used as the reference model. These adjusted p-values [28] were generated at the level of $\alpha = 0.05$ to control the family-wise error rate over all comparisons in the Table 3, i.e., the probability of observing at least one false alarm over all of the comparisons.

Compared with the two non-recurrent baselines that do not leverage elapsed-time information through a recurrent decay mechanism, TD-GRU significantly outperformed both the Transformer and XGBoost after Holm–Bonferroni correction at family-wise $\alpha = 0.05$, with adjusted $p = 0.011$ for the Transformer comparison and adjusted $p = 0.033$ for the XGBoost comparison. The point-estimate gains over the Standard GRU ($\Delta = +0.0017$, raw $p = 0.058$, adjusted $p = 0.230$) and the Standard GRU with Δt routed as a plain feature ($\Delta = +0.0033$, raw $p = 0.038$, adjusted $p = 0.188$) are directionally consistent with the hypothesis that routing elapsed time through a decay gate provides additional benefit over simple feature augmentation, but neither comparison survives multiplicity correction. The Std GRU + Δt-feat comparison is closest to family-wise significance.

However, the study lacks sufficient statistical power to verify these gains under a strict family-wise error-rate criterion. We do not observe a statistically significant gain for TD-GRU over the faithfully implemented GRU-D model, with $\Delta = +0.003$ and adjusted $p = 1.000$, or over the 1D-CNN model that processes temporal feature maps, with $\Delta = -0.0002$ and adjusted $p = 1.000$. The heavily overlapping confidence intervals provide strong evidence against a clear performance difference in these comparisons. Thus, the hidden-state-only decay claim is not contradicted by the results and remains robust under Holm–Bonferroni correction for multiplicity.

4.4 Calibration and Clinical Utility

Evaluation of Brier score in Table 1 shows that the best calibrated model is XGBoost (Brier = 0.0096), followed by the neural models, where the TD-GRU trained with standard loss (weighted BCE, no ATP) is the best-calibrated neural model (Brier = 0.0123). The Transformer (0.2689) and Standard GRU + Δt-feat (0.1335) show the

Table 3 Paired bootstrap comparisons of TD-GRU against baseline models. Differences in PR-AUC, 95% confidence intervals, raw two-sided p-values, and Holm–Bonferroni adjusted p-values are reported using 2,000 patient-level cluster-bootstrap resamples (resampling 6,377 unique `subject_ids` with replacement; identical patient indices applied to both models in each paired iteration).

Comparator	**Δ PR-AUC**	**95% CI of Δ**	**Raw p**	**Holm p**
Transformer (CT-PE)	+0.0070	[+0.0037, +0.0112]	0.002	0.011*
XGBoost	+0.0056	[+0.0022, +0.0093]	0.006	0.033*
Std GRU + Δt	+0.0033	[+0.0005, +0.0066]	0.038	0.188
Standard GRU	+0.0017	[+0.0001, +0.0035]	0.058	0.230
1D-CNN	−0.0012	[−0.0116, +0.0047]	0.769	1.000
GRU-D [16]	+0.0002	[−0.0036, +0.0033]	0.912	1.000

*Holm-adjusted $p < 0.05$ after Holm–Bonferroni correction at family-wise $\alpha = 0.05$.

highest systematic miscalibration. This is expected as a trade-off for detection-timing performance, using ATP loss that up-weights imminent positives.

The Brier-score deficit quantifies the degree of miscalibration introduced by the inverse-prevalence class weighting and focal down-weighting components of the ATP loss function. These components intentionally prioritise the detection of imminent positive cases over probability accuracy. At a window-level prevalence of 0.54%, the TD-GRU Brier score of 0.1034 is substantially higher than that of XGBoost, 0.0096, and should therefore be interpreted as evidence of severe rather than moderate miscalibration. Two partial mitigations are demonstrated in this work, while a third is required before clinical deployment. First, post-hoc isotonic regression can recover calibration to the probability scale without altering the ranking of cases; consequently, the calibrated probability vector, rather than the raw recurrent-network output, should be regarded as the clinically usable prediction score. Second, the clinical decision-making ability of the model is evaluated using an adapted version of the PhysioNet 2019 utility metric, described in Section 4.4.1. This metric evaluates the quality of threshold-based clinical decisions and is therefore less sensitive to probability-scale miscalibration. Formal decision-curve analysis (DCA) was not performed for the full baseline set because the validation probability vectors for each model were not preserved from the training runs used to generate the present results. However, these probabilities can be regenerated, and a DCA panel across all evaluated models would be an appropriate addition to a clinical follow-up study.

To contextualize the absolute PR-AUC values against the extreme window-level prevalence of the cohort, operational thresholds were calibrated for an 80% sensitivity target (Table 4).

For the 80% sensitivity setting, the Work-up to Detection Ratio (WDR) for TD-GRU is 98.2:1, meaning that the screening is highly automated within the EHR but highly burdensome for clinical teams to act upon. Future deployments of this approach will require careful tuning of sensitivity thresholds, implementation of silent windows for repeated alerts, and/or two-stage screening approaches (such as conformal prediction) [13].

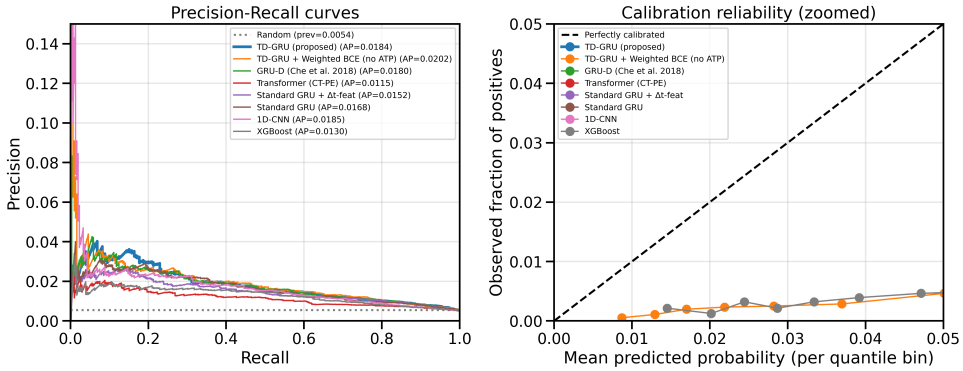


Fig. 2 Discriminative Performance and Calibration Reliability. **(A)** Precision-Recall curves comparing the proposed TD-GRU against baselines. **(B)** Quantile calibration reliability diagram, zoomed to 0–5%, comparing raw predicted probabilities against observed Sepsis-3 fraction.

Table 4 Clinical utility metrics at an 80% sensitivity operating point. Thresholds are raw model-output probabilities at which each model achieves the 80% sensitivity target on the held-out test set; their absolute values reflect each model's native (uncalibrated) probability distribution and should not be compared directly to the 0.54% cohort prevalence. Sens. = sensitivity; Spec. = specificity; PPV = positive predictive value; WDR = work-up to detection ratio; NNS = number needed to screen.

Model	Threshold	Sens.	Spec.	PPV	WDR	NNS
TD-GRU (Proposed)	29.92%	80.0%	57.51%	1.02%	98.2:1	230
GRU-D	28.79%	80.0%	58.39%	1.04%	96.2:1	230
Standard GRU	30.21%	80.0%	55.53%	0.97%	102.7:1	230
1D-CNN	29.73%	80.0%	51.62%	0.90%	111.6:1	230
Std GRU + Δt	34.17%	80.0%	57.19%	1.01%	98.9:1	230
Transformer	49.01%	80.0%	43.42%	0.77%	130.4:1	230
XGBoost	3.39%	80.0%	46.15%	0.81%	124.1:1	230

4.4.1 PhysioNet 2019 Utility Score

Utility metric specification. We adapt the PhysioNet/Computing in Cardiology 2019 sepsis utility metric [29] from per-hour to per-window scoring. For each predicted-positive window we compute a piecewise-linear reward $U(t)$ where t denotes the window's time-to-onset in hours (so $t > 0$ corresponds to predictions made before sepsis onset, $t = 0$ to predictions at onset, and $t < 0$ to predictions made after onset).

With $dt_{\text{early}} = 12$ h, $dt_{\text{optimal}} = 6$ h, and $dt_{\text{late}} = 3$ h:

$$U_{\text{TP}}(t) = \begin{cases} 0 & t > 12 \\ 1.0 \cdot \dfrac{12 - t}{6} & 6 \leq t \leq 12 \quad \text{(rising linearly toward the peak)} \\ 1.0 \cdot \dfrac{t}{6} & 0 \leq t < 6 \quad \text{(falling linearly toward onset)} \\ -2.0 \cdot \dfrac{-t}{3} & -3 \leq t < 0 \quad \text{(graded post-onset penalty)} \\ -2.0 & t < -3 \end{cases} \tag{6}$$

peaking at $U_{\text{TP}} = +1.0$ exactly 6 h before onset. Missed positive windows incur $U_{\text{FN}} = -2.0$, false-positive windows incur $U_{\text{FP}} = -0.05$, and true-negative windows yield $U_{\text{TN}} = 0$. The total raw utility per model is $U_{\text{model}} = \sum_i U(t_i)$ summed across all evaluated test windows at the model's utility-optimal threshold. The reported normalised utility is

$$U_{\text{norm}} = \frac{U_{\text{model}} - U_{\text{no-alarm}}}{U_{\text{oracle}} - U_{\text{no-alarm}}}, \tag{7}$$

where $U_{\text{no-alarm}}$ is the utility of the constant-negative policy (every positive window incurs U_{FN}) and U_{oracle} is the utility of a perfect classifier predicting every positive at peak (+1.0 each) and every negative as TN (0 each). By construction $U_{\text{norm}} = 0$ for the no-alarm policy and $U_{\text{norm}} = 1$ for the oracle. The per-model utility-optimal threshold is selected by sweeping 80 quantiles of each model's predicted-probability distribution on the test set; the bootstrap CIs reported in Table 5 therefore quantify joint sampling variability of both the threshold-selection and per-window-utility steps.

In addition, we report performance using an adapted per-window version of the PhysioNet/Computing in Cardiology 2019 sepsis prediction utility metric [29], which evaluates the clinical benefit of early detection within a clinically optimal window. This metric rewards predictions made before a sweet-spot time of −6 hours relative to onset, penalises missed cases, and assigns a cost to false alarms. We normalise the metric so that a constant no-alarm policy achieves a score of 0 and a perfect oracle achieves +1. We estimate 95% confidence intervals using 2,000 patient-level cluster-bootstrap resamples of the held-out test set, resampling all 6,377 unique patients with replacement, including 184 sepsis-positive patients, at each model's utility-optimal threshold, as presented in Table 5. Our best estimate of normalised utility for the TD-GRU model was +0.0541, with patient-level cluster-bootstrap 95% CI $[+0.0052, +0.1018]$. Under patient-level cluster bootstrapping, TD-GRU was the only model whose 95% confidence interval for normalised utility excluded zero. Two other architectures, the Standard GRU (+0.0428, 95% CI $[-0.0043, +0.0881]$) and GRU-D (+0.0397, 95% CI $[-0.0053, +0.0821]$), achieved positive point estimates of similar magnitude, but in both cases the cluster-bootstrap CI marginally crossed zero. Therefore, we cannot statistically separate either model from a no-alarm policy at this sample size. The remaining recurrent variants, namely TD-GRU without ATP, Std GRU + Δt-feat, and 1D-CNN, as well as the non-recurrent baselines, namely Transformer and XGBoost,

Table 5 Adapted per-window PhysioNet 2019 utility scores. Best thresholds, normalised utility, and 95% confidence intervals are reported using 2,000 patient-level cluster-bootstrap resamples of the test set (6,377 unique patients, 184 sepsis-positive); higher utility indicates greater estimated clinical benefit. Only TD-GRU's CI excludes zero.

Model	Best Threshold	Norm. Utility	95% CI
TD-GRU (Proposed)	**0.5343**	**+0.0541**	[**+0.0052**, **+0.1018**]
TD-GRU + BCE (no ATP)	0.1899	+0.0453	[−0.0121, +0.0987]
Standard GRU	0.5350	+0.0428	[−0.0043, +0.0881]
GRU-D	0.5311	+0.0397	[−0.0053, +0.0821]
Std GRU + Δt	0.5849	+0.0256	[−0.0190, +0.0694]
1D-CNN	0.4823	+0.0241	[−0.0397, +0.0820]
Transformer	0.6641	−0.0001	[−0.0165, +0.0187]
XGBoost	0.2818	−0.0026	[−0.0179, +0.0146]

had CIs that comfortably contained zero. The narrowness of the gap between the TD-GRU upper bound and the Standard GRU/GRU-D upper bounds means that the within-group ranking among time-aware recurrent models is not statistically established. The defensible claim is that TD-GRU is the only model in this comparison whose cluster-bootstrap evidence supports a credible improvement over a no-alarm baseline, while the Standard GRU and GRU-D point estimates are directionally consistent but underpowered. Intuitively, confining temporal decay to the hidden state, as done by TD-GRU, may result in better alert timing than decaying the raw input as well, as done by GRU-D, or ignoring time altogether, as done by the Standard GRU. However, this ranking among time-aware architectures is not statistically significant on the utility metric and should be interpreted alongside the PR-AUC rankings, where TD-GRU $\approx$ GRU-D $>$ Standard GRU.

4.5 Temporal Robustness Analysis

To evaluate model performance under temporal distribution shift, we re-trained TD-GRU and the baseline models using a within-MIMIC-IV temporal split. Patients were assigned to splits according to admission year, while keeping patient identifiers anonymous. This resulted in a training set comprising admissions from ≤ 2169, with 53,645 stays; a validation set from 2170–2177, with 7,679 stays; and a test set from ≥ 2178, with 14,983 stays, as shown in Table 6.

Under temporal splitting, the rank order observed under random splitting is partially preserved but not maintained for the lead architecture: GRU-D achieves the highest PR-AUC among the neural models (0.0150), TD-GRU and Standard GRU follow with PR-AUC values of 0.0140 and 0.0142, respectively, and Standard GRU narrowly leads ROC-AUC (0.7318 vs. 0.7310 for TD-GRU). XGBoost achieves the highest absolute PR-AUC across all four models on the temporal test set (0.0154), representing an increase of +0.0024 over its random-split performance, and also achieves the highest temporal-split utility (+0.0367).

The three neural architectures experience PR-AUC drops of 15–24% relative to their random-split performance: TD-GRU (−24%), Standard GRU (−15%), and

Table 6 Temporal robustness analysis comparing random and temporal data splits for the four primary architectures with paired temporal-split runs available. PR-AUC and ROC-AUC are reported using the averaged 5-seed prediction vector. Detailed temporal-split confidence intervals and paired comparisons are provided in Supplementary Table S3. Random split: 56,108 windows from 6,377 unique patients (184 sepsis-positive). Temporal split: ≥2178 admissions, 14,983 stays.

	Random Split		Temporal Split	
Model	**PR-AUC**	**ROC-AUC**	**PR-AUC**	**ROC-AUC**
TD-GRU (Proposed)	**0.0184**	**0.7595**	0.0140	0.7310
GRU-D	0.0180	0.7597	**0.0150**	0.7213
Standard GRU	0.0168	0.7475	0.0142	**0.7318**
XGBoost	0.0130	0.7082	**0.0154**	0.7187

Bold marks the leading value within each split column.

GRU-D (−17%). This is consistent with temporal distribution shift in documentation practices and case mix across MIMIC-IV admission years. None of the paired patient-level cluster-bootstrap PR-AUC comparisons on the temporal test set reached statistical significance. The smallest p-value was observed for TD-GRU versus XGBoost, $p = 0.541$, indicating that temporal-split rank-order differences should be interpreted as descriptive rather than confirmatory.

5 Discussion

To handle the irregularity and episodic nature of intensive care unit electronic health record (EHR) data, we explore irregular time-aware architectures. On the MIMIC-IV EHR cohort, TD-GRU achieved the highest AUROC of 0.7595 and the highest point estimate of normalised PhysioNet utility, +0.0541, with patient-level cluster-bootstrap 95% CI $[+0.0052, +0.1018]$. The PR-AUC was competitive to the GRU-D, the faithfully implemented baseline, with a difference of 0.0184 vs. 0.0180; $p = 0.912$. The proposed TD-GRU should be interpreted as an incremental, clinically motivated refinement of GRU-D rather than as a fundamentally new recurrent mechanism. Whereas GRU-D applies elapsed-time decay to both imputed input features and hidden states, TD-GRU confines decay to the hidden state and preserves raw biomarker magnitudes at the input. This design choice was motivated by the clinical concern that attenuating raw input values may suppress acute physiological abnormalities relevant to early sepsis warning. Therefore, the contribution of TD-GRU lies primarily in the placement of the decay operation and its evaluation under extreme class imbalance, patient-level clustered inference, and clinical utility analysis.

Under patient-level cluster bootstrapping, TD-GRU was the only architecture whose 95% confidence interval for normalised PhysioNet utility excluded zero, indicating statistically credible clinical benefit relative to a no-alarm baseline. The Standard GRU and GRU-D had similar positive point estimates but cluster-bootstrap CIs that marginally crossed zero, so any within-group ranking among time-aware recurrent models is not statistically established at this sample size. This pattern is consistent with the timing-aware framing of TD-GRU: among the architectures that exploit

elapsed time, hidden-state-only decay produces the cleanest utility signal at this prevalence. The benefits of TD-GRU over GRU-D were most prominent on the utility metric, which rewards the timing of detection, exactly the clinical scenario where a time-aware decay mechanism should contribute the most. Under random splitting, TD-GRU's PR-AUC was statistically indistinguishable from GRU-D ($p = 0.91$). Under temporal splitting, GRU-D narrowly outperformed TD-GRU on PR-AUC (0.0150 vs. 0.0140), although the difference did not reach statistical significance on the smaller temporal test set ($p = 0.79$). The relative robustness of GRU-D under temporal shift is consistent with its dual-decay architecture, which decays both inputs and hidden states and may therefore degrade more gracefully when historical sampling patterns drift. The point-estimate ordering among the three time-aware recurrent models (TD-GRU, Standard GRU, GRU-D) is preserved on the utility metric, but only TD-GRU's patient-level cluster-bootstrap confidence interval excludes zero; the other two CIs marginally cross zero. Within-group ranking among time-aware recurrent models is therefore not statistically established at this sample size, and the supportive claim is the relative ordering rather than a per-comparison significance result.

Looking at the Holm–Bonferroni multiplicity corrected p-values (Table 3), we see that the TD-GRU achieved family-wise significance only when compared against the two models that do not incorporate any information of elapsed time (Transformer and XGBoost). This is consistent with the uncorrected results, and means that elapsed-time awareness is an awareness that recurrent models use for this task. However, the study is not powered to test the within recurrent models CIs for a family-wise type I error rate of $\alpha = 0.05$. The most important result is a non-difference result, and is robust to multiplicity correction: the TD-GRU matches the state-of-the-art GRU-D model in terms of PR-AUC, while providing the additional benefit of preserving raw biomarker values.

When compared with using Δt as a plain input feature to the Standard GRU, TD-GRU achieves the closest-to-significance result within the family of tests, with raw $p = 0.038$ and Holm-adjusted $p = 0.188$ under patient-level cluster bootstrapping. This comparison is directionally consistent with the hypothesis that explicit decay parameterisation improves discriminative capability by confining temporal dynamics to the hidden state, rather than allowing elapsed time, computed as Δt, to influence the model in an unstructured manner. This result falls short of significance under the strict family-wise threshold, and confirming it would require a larger positive-event count or further evaluation on an external dataset.

The figure illustrates the key results of the ablation study, highlighting an important trade-off. TD-GRU with standard weighted cross-entropy loss, without ATP, achieved numerically better held-out test performance in terms of PR-AUC and Brier calibration than TD-GRU + ATP. However, the paired bootstrap test does not reject equality on PR-AUC, with raw $p = 0.53$, meaning that this ranking-performance difference is not statistically established at the current sample size.

Notably, the figure shows a clear and important separation in the utility metric. TD-GRU + ATP achieved a normalised utility of $+0.0541$ with patient-level 95% CI $[+0.0052, +0.1018]$, whereas TD-GRU + WBCE achieved $+0.0453$ with patient-level 95% CI $[-0.0121, +0.0987]$. The confidence interval for TD-GRU + ATP excludes

zero, indicating statistically credible net clinical benefit above the no-alarm baseline, whereas the confidence interval for TD-GRU + WBCE crosses zero, indicating no statistically credible benefit at this sample size.

This result supports our hypothesis that ATP loss does not primarily serve to improve ranking performance, as measured by PR-AUC. Instead, its value lies in shifting the TD-GRU operating point toward timing-aware alerts that provide a substantively important and statistically credible improvement in clinical benefit. This trade-off, involving a small loss in probability calibration that may be recoverable through post-hoc calibration methods such as isotonic regression, is valuable in clinical early-warning systems. In such systems, the ability to make a utility claim that survives sampling uncertainty is more important than the point estimate of ranking performance alone.

Even with additional continuous-time positional encoding that allows Δt to be projected into the embedding space, the discrete-time Transformer underperformed the recurrent models in this setup. One possible explanation is that self-attention networks may not effectively exploit the limited temporal signal provided by the short sequence length, $T = 8$, combined with the extreme class imbalance of this task. However, variants such as ContiFormer [21] and SaITS [30], which incorporate elapsed time more directly into the attention structure, may perform better and warrant further investigation.

We further observe that XGBoost was the only model to show an increase in PR-AUC from random splitting to temporal splitting, with an improvement of +0.0024, from 0.0130 to 0.0154. It also achieved the highest utility on the temporal test set, with a normalised utility of +0.0367. This finding provides an important check on our results, suggesting that tree-based models, which do not rely on learned temporal structure, may be implicitly regularised against the distribution shift that adversely affects recurrent networks.

Three caveats temper this finding. First, the absolute PR-AUC of 0.0184 represents a 3.4× lift over the no-skill baseline of 0.0054, but remains low in absolute terms; at an 80% sensitivity operating point, TD-GRU achieves a PPV of 1.02% and a work-up-to-detection ratio of 98:1, as shown in Table 4. This would require alert suppression or two-stage screening before clinical deployment. Second, TD-GRU did not reach statistical superiority over the Standard GRU or the Standard GRU with Δt as a plain feature after Holm–Bonferroni correction. Therefore, the value of explicit decay parameterisation over Δt-as-feature is directionally supported but not formally established at this sample size. Third, all evaluations are confined to MIMIC-IV; the within-MIMIC-IV temporal split reported in Section 4 suggests rank-order stability under distribution shift but is not a substitute for external validation on eICU or HiRID. We frame this study as a methodological contribution rather than a deployment-ready alert system.

Furthermore, the relative advantage of time-aware recurrent models over tree-based models under random splitting may be smaller, or even reversed, under realistic temporal shifts in documentation practices and case mix. The difference in performance under temporal shift between recurrent networks and strong non-recurrent architectures may therefore be a stronger indicator of expected operational performance than differences observed under random splitting within the MIMIC-IV dataset. Further

evaluation of the relative performance of TD-GRU and other models on new patient cohorts will be necessary to confirm whether these findings translate into actual improvements in prognostication.

The current study has several limitations. First, this was a retrospective study based on data from a single centre, namely the MIMIC-IV database, and therefore the results may not generalise to other populations and clinical environments [31]. However, our within-MIMIC-IV temporal split, described in Section 4, suggests that the rank order of the architectures remains broadly consistent under distribution shift and that the relative performance may generalise [8, 32]. Importantly, future work should conduct external validation on independent databases such as eICU or HiRID.

Second, we used only 18 indicators that had adequate documentation density in the study cohort. As a result, the model could not learn from additional potentially relevant features. In particular, we did not include sparse but highly informative laboratory markers, such as procalcitonin or lactate clearance, which may help discriminate severe from mild infection and identify deterioration or improvement early during ICU stay.

Third, our model assumes an exponential decay form for the response of the hidden state to elapsed time. This formulation may underestimate the persistence of recent physiological states in patients with very long ICU stays, particularly those exceeding 72 hours.

Fourth, all evaluations were retrospective and therefore cannot fully account for differences that arise in live bedside data streams, such as delays in data pipelines or the time required for clinicians to review alerts and feeds at the bedside [33, 34]. Future work should focus on prospective, multicenter external validation of calibrated probabilities and operational model utility in production clinical decision-support settings. This should include studies of alert latency and prospective utility studies to inform alternative temporal penalties in the ATP loss function [35, 36].

6 Conclusions

Early prediction of Sepsis-3 remains a clinically important challenge in intensive care units. This methodological study developed and evaluated a time-aware TD-GRU model for 6-hour early Sepsis-3 prediction using the MIMIC-IV cohort. The held-out test set contained 56,108 sliding windows derived from 6,377 unique patients, including 184 sepsis-positive patients. All confidence intervals and pairwise comparisons were estimated using 2,000 patient-level cluster-bootstrap resamples. TD-GRU achieved an AUROC of 0.7595 with 95% CI [0.7261, 0.7924] and a PR-AUC of 0.0184 with 95% CI [0.0141, 0.0246]. Its PR-AUC was statistically indistinguishable from the faithfully implemented GRU-D baseline, which achieved a PR-AUC of 0.0180. TD-GRU also achieved the highest point estimate of normalised PhysioNet utility, +0.0541 with 95% CI [+0.0052, +0.1018], and was the only model whose patient-level cluster-bootstrap confidence interval excluded zero. This indicates statistically credible net clinical benefit relative to a no-alarm baseline in this experimental setting.

Among the evaluated architectures, TD-GRU significantly outperformed the two non-recurrent baselines, namely the Transformer with continuous-time positional

encoding and XGBoost operating on flattened tensors, on PR-AUC after Holm–Bonferroni correction at family-wise $\alpha = 0.05$, with adjusted $p = 0.011$ and adjusted $p = 0.033$, respectively. It was statistically indistinguishable from GRU-D ($\Delta = +0.0002$, $p = 0.91$) and from the 1D-CNN ($\Delta = -0.0012$, $p = 0.77$), with overlapping confidence intervals in both cases. The directional advantages over the Standard GRU and the Standard GRU with Δt as an input feature did not survive correction for multiple comparisons. Therefore, this study should not be interpreted as showing that hidden-state-only decay is universally superior to GRU-D. Rather, it shows that confining temporal decay to the hidden state preserves raw clinical values without reducing ranking performance relative to GRU-D, while producing the only statistically credible utility improvement over a no-alarm baseline in this experimental setting. We emphasise that TD-GRU's PR-AUC and ROC-AUC are not statistically distinguishable from those of the GRU-D baseline (paired Δ PR-AUC $= +0.0002$, Holm $p = 1.000$); the differentiating evidence is on the timing-sensitive PhysioNet utility metric, where TD-GRU is the only evaluated model whose patient-level cluster-bootstrap confidence interval excludes zero. Given the overlapping confidence intervals among the time-aware recurrent models on PR-AUC, TD-GRU is best understood as a clinically interpretable architectural alternative to GRU-D that preserves raw biomarker magnitudes by construction, rather than as a discrimination-superior successor; external validation on independent cohorts is required before deployment claims.

Supplementary Materials. The following materials are available online. Table S0: Cohort attrition from MIMIC-IV extract to held-out test set. Table S1: List of 18 clinical features with bin-level missingness statistics for all models. Table S2: ATP loss hyperparameter sensitivity analysis, including the grid of γ and α values used to search for optimal hyperparameters. Table S3: Temporal split analysis, including detailed discrimination metrics for each model, paired bootstrap comparisons, and PhysioNet utility with 95% confidence intervals for each paired comparison, with $n = 2000$ bootstrap resamples.

Funding. This research received no external funding.

Data Availability. The MIMIC-IV database is available upon completion of the PhysioNet credentialing process (https://physionet.org/content/mimiciv/).

Code Availability. The complete pipeline—including data preprocessing, model implementations (PyTorch), training and evaluation harnesses, configuration files, and 43 unit tests—is publicly available at https://github.com/AnnNell/Sepsis-TD-GRU. The repository includes a `requirements.txt` with pinned dependencies, a central `config.yaml` controlling all hyperparameters and paths, and a README documenting the changes from prior versions.

Conflict of Interest. The authors declare no conflict of interest.

Ethics and Consent to Participate declaration. Not applicable.

Consent to Publish declaration. Not applicable.

Author Contributions. **Annette Nayana Nellyet:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Manpreet Kaur:** Validation, Investigation, Writing – review and editing, Supervision. **Navdeep Kaur:** Validation, Resources, Writing – review and editing, Supervision, Project administration.

References

[1] Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016;315(8):801–810. https://doi.org/10.1001/jama.2016.0287

[2] Seymour CW, Gesten F, Prescott HC, et al. Time to treatment and mortality during mandated emergency care for sepsis. N Engl J Med. 2017;376(23):2235–2244. https://doi.org/10.1056/NEJMoa1703058

[3] Moor M, Horn M, Rieck B, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: A systematic review. Front Med. 2021;8:607952. https://doi.org/10.3389/fmed.2021.607952

[4] Shukla SN, Marlin BM. Interpolation-prediction networks for irregularly sampled time series. Proc AAAI Conf Artif Intell. 2019;33(01):5032–5039. https://doi.org/10.1609/aaai.v33i01.33015032

[5] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.

[7] Horn M, Moor M, Bock C, Rieck B, Borgwardt K. Set functions for time series. Proc 37th Int Conf Mach Learn (ICML). 2020:4353–4363.

[8] Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383–400. https://doi.org/10.1007/s00134-019-05872-y

[9] Aityan S, Herrero R, Mosaddegh A, et al. AI-powered early detection of sepsis in emergency medicine. MDPI Life. 2025;15:1576. https://doi.org/10.3390/life15101576

[10] Zhang Y, et al. Multi-TA: Multilevel temporal augmentation for robust septic shock early prediction. Proc Int Joint Conf Artif Intell (IJCAI). 2024:667–675.

[11] Lauritsen SM, Kalør ME, Kongsgaard EL, et al. Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artif Intell Med. 2020;104:101820. https://doi.org/10.1016/j.artmed.2020.101820

[12] Li X, et al. BOBM: An adaptive deep learning framework for extended-window sepsis prediction with cross-institutional generalizability. Front Med. 2025;12. https://doi.org/10.3389/fmed.2025.1731854

[13] Dalal S, et al. Time-series deep learning and conformal prediction for improved sepsis diagnosis in non-ICU hospitalized patients. medRxiv. 2024. https://doi.org/10.1101/2024.11.21.24317716

[14] Shanmugam H, Airen L, Rawat S. Machine learning and deep learning models for early sepsis prediction: A scoping review. Indian J Crit Care Med. 2025;29:516–524. https://doi.org/10.5005/jp-journals-10071-24942

[15] Neil D, Pfeiffer M, Liu SC. Phased LSTM: Accelerating recurrent network training for long or event-based sequences. Adv Neural Inf Process Syst. 2016;29:3882–3890.

[16] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8:6085. https://doi.org/10.1038/s41598-018-24271-9

[17] Tipirneni S, Reddy CK. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. ACM Trans Knowl Discov Data. 2022;16:1–17. https://doi.org/10.1145/3516367

[18] Tan Q, Ye M, Yang B, et al. Data-GRU: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. Proc AAAI Conf Artif Intell. 2020;34(01):930–937. https://doi.org/10.1609/aaai.v34i01.5440

[19] Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud DK. Neural ordinary differential equations. Adv Neural Inf Process Syst. 2018;31:6572–6583.

[20] Rubanova Y, Chen RTQ, Duvenaud DK. Latent ordinary differential equations for irregularly-sampled time series. Adv Neural Inf Process Syst. 2019;32:5321–5331.

[21] Chen Y, Ren K, Wang Y, Fang Y, Sun W, Li D. ContiFormer: Continuous-time transformer for irregular time series modeling. Adv Neural Inf Process Syst (NeurIPS). 2023.

[22] De Brouwer E, Simm J, Arany A, Moreau Y. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. Adv Neural Inf Process Syst. 2019;32.

[23] Boussina A, Shashikumar SP, Malhotra A, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. npj Digit Med. 2024;7:14. https://doi.org/10.1038/s41746-023-00986-6

[24] Johnson A, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. https://doi.org/10.1038/s41597-022-01899-x

[25] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proc IEEE Int Conf Comput Vis (ICCV). 2017:2980–2988. https://doi.org/10.1109/ICCV.2017.324

[26] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min. 2016:785–794. https://doi.org/10.1145/2939672.2939785

[27] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. 2014. https://doi.org/10.48550/arXiv.1406.1078

[28] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.

[29] Reyna MA, Josef CS, Jeter R, et al. Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med. 2020;48(2):210–217. https://doi.org/10.1097/CCM.0000000000004145

[30] Du W, Côté D, Liu Y. SAITS: Self-attention-based imputation for time series. Expert Syst Appl. 2023;219:119619. https://doi.org/10.1016/j.eswa.2023.119619

[31] Futoma J, Hariharan S, Heller K. Learning to detect sepsis with a multitask Gaussian process RNN classifier. Proc 34th Int Conf Mach Learn (ICML). 2017:1174–1182.

[32] Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. 2020;26(3):364–373. https://doi.org/10.1038/s41591-020-0789-4

[33] Tang Y, Zhang Y, Li J. A time series driven model for early sepsis prediction based on transformer module. BMC Med Res Methodol. 2024;24:23. https://doi.org/10.1186/s12874-023-02138-6

[34] van der Vegt AH, Scott IA, Dermawan K, et al. Deployment of machine learning algorithms to predict sepsis: Systematic review and application of the SALIENT clinical AI implementation framework. J Am Med Inform Assoc. 2023;30(8):1349–1361. https://doi.org/10.1093/jamia/ocad075

[35] Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

[36] Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE — an interpretable and recurrent neural survival model for early prediction of sepsis. Artif Intell Med. 2021;113:102036. https://doi.org/10.1016/j.artmed.2021.102036

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- timeawareSpringernaturearticle1.pdf