Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.087214

ARTICLE

Sparse Physio-Attention: A Computationally Efficient and Clinically Interpretable Framework for ICU Time-Series Analysis

Hashim Ali*

School of Computing and Artificial Intelligence, Nazarbayev University, Astana, Kazakhstan

*Corresponding Author: Hashim Ali. Email: hashim.ali@nu.edu.kz

Received: 12 June 2026; Accepted: 20 July 2026; Published: 13 August 2026

ABSTRACT: Intensive care unit (ICU) time series are irregular, incomplete, and computationally demanding to model at high temporal resolution. Dense Transformer attention captures long-range dependencies but evaluates all pairwise interactions, including many stable or clinically weak measurements. This study presents Sparse Physio-Attention, a physiology-guided Transformer that retains critical-range violations, patient-relative deviations, informative missingness patterns, and task-relevant variables before sparse attention is computed. Dynamic routing subsequently removes weak attention edges, and a late-fusion adapter incorporates static electronic health record context. The analysis included 25,368 eligible MIMIC-IV ICU stays, of which 2740 were sepsis positive. On the held-out test set, the proposed model achieved an AUROC of 0.892 and an AUPRC of 0.835 at the 24-h prediction horizon, with an expected calibration error of 0.031. Relative to dense attention, it reduced floating-point operations by approximately 72%, peak GPU memory from 3.2 to 1.8 GB, and 24-h-window latency from 670 ± 24 to 235 ± 11 ms. Attention-guideline agreement reached 0.826 at 24 h, indicating alignment with prespecified deterioration rules; this result reflects clinical plausibility rather than causal explanation. The evidence is limited to internal MIMIC-IV validation, and the rule-based retention thresholds may require recalibration across hospitals and patient subgroups. External and prospective validation is therefore required before clinical deployment.

KEYWORDS: Intensive care unit; clinical time series; sparse attention; transformer; explainable artificial intelligence; clinical decision support; sepsis prediction

1 Introduction

The intensive care unit (ICU) is a data-rich clinical environment in which physiological measurements, laboratory results, medication records, ventilator settings, and demographic variables are collected repeatedly during patient care. These data streams are valuable for early detection of deterioration, including sepsis, respiratory failure, acute kidney injury, prolonged length of stay, and mortality. However, ICU data also present substantial modeling challenges because measurements are irregularly sampled, frequently missing, affected by clinical workflows, and strongly dependent on patient context [1–4].

Deep learning methods have increasingly been used for ICU prediction because they can model nonlinear interactions and temporal dependencies. Recurrent neural networks, including long short-term memory (LSTM) models, can represent sequential patterns but are difficult to parallelize and may struggle with long-range dependencies [5,6]. Transformer architectures use self-attention to model relationships among sequence elements and have shown strong performance in clinical time-series tasks [7–9]. Nevertheless,

dense self-attention has quadratic complexity with respect to sequence length, which is undesirable when high-frequency ICU data must be processed under near-real-time constraints.

The second challenge concerns interpretability and workflow fit. In clinical decision support systems (CDSSs), model predictions are useful only when clinicians can judge whether an alert is timely, calibrated, and medically plausible. ICU-focused evaluations indicate that explanation quality should be assessed by its effect on clinician decision-making and by whether the highlighted variables are clinically plausible and free from evident bias [10,11]. Attention weights can provide useful inspection signals, but they should not be interpreted as complete causal explanations. In ICU modeling, a clinically readable attention mechanism should highlight physiologically meaningful events, such as falling blood pressure, rising lactate, worsening oxygen saturation, increasing respiratory rate, or deteriorating renal function.

Sparse Physio-Attention addresses these limitations by integrating physiological priors into sparse attention computation. In this paper, physiological prior refers to domain knowledge about critical ranges, abnormal temporal changes, organ-system-specific variables, and patient-contextual baselines. The prior is not treated as a rigid substitute for learning; rather, it defines a safety-aware retention mechanism that preserves clinically important time-variable pairs before learned sparse attention is applied. This distinction is important because critically ill patients may have abnormal baselines, and the same measured value can carry different meaning depending on age, comorbidities, admission diagnosis, and early ICU trajectory.

The main methodological contributions are as follows.

- First, a leakage-aware physiology-guided temporal and feature pruning strategy is proposed for ICU time-series modeling.
- Second, a hybrid sparse-dynamic attention mechanism is developed to combine clinical retention rules with learned routing of attention connections.
- Third, a cross-modal adapter is introduced to integrate static electronic health record (EHR) variables with sparse temporal representations.
- Fourth, the framework is evaluated against recurrent, convolutional, Transformer, sparse-attention, and clinical sequence baselines using predictive, computational, calibration, operating-point, and interpretability metrics.

The remainder of this paper is organized as follows. Section 2 reviews related work and state-of-the-art methods. Section 3 presents the proposed Sparse Physio-Attention framework, including physiological masking, sparse attention, and cross-modal EHR fusion. Section 4 describes the datasets, baselines, metrics, and implementation details. Section 5 reports computational, predictive, calibration, interpretability, and ablation results. Section 6 discusses clinical relevance, limitations, ethical considerations, and future directions. Section 7 concludes the paper.

2 Related Work

2.1 Clinical Time-Series Prediction in the ICU

ICU prediction has evolved from conventional scoring systems and statistical models toward data-driven machine learning. Public databases such as MIMIC-IV and eICU-CRD have enabled reproducible research on mortality, sepsis, readmission, and length-of-stay prediction [3,4,12]. Early models relied on manually engineered features, autoregressive structures, or gradient-based learners. Intensive care unit (ICU) prognostic models must be designed to account for irregularly sampled data, informative patterns of missingness, pronounced class imbalance, and rapidly evolving physiological states. Gated Recurrent Unit with Decay (GRU-D) represents missingness through observation masks and elapsed-time decay, whereas ICU-specific temporal models use recurrent, convolutional, or attention-based encoders to learn

deterioration trajectories [6,8]. For sepsis, Tang et al. [9] combined convolutional or recurrent feature extraction with a Transformer module, and Xu et al. [13] used clinical time-series forecasting to support early Sepsis-3 prediction. These models establish the value of temporal attention for early warning, but they do not explicitly preserve clinically critical observations before attention computation.

Interpretability has also been investigated directly in early sepsis modeling. Multitask Gaussian Process and Attention-based Temporal Convolutional Network (MGP-AttTCN) [14] combines Gaussian-process handling of irregular observations with temporal attention and reports clinically inspectable sepsis predictions. Temporal convolution-attention models likewise identify physiological features associated with early sepsis while retaining a supervised prediction objective [15]. Sparse Physio-Attention differs from these approaches by using clinical retention rules to reduce the token set before a sparse transformer encoder is applied.

2.2 Sparse and Efficient Attention

Standard self-attention computes pairwise interactions across all sequence elements. For a sequence of T tokens, dense self-attention requires $O(T^2)$ pairwise score computation. Recent sepsis-specific work has reduced this burden without relying solely on generic long-sequence architectures. The KA-Transformer replaces conventional dot-product attention with kernel attention to reduce parameter burden and model nonlinear temporal patterns in early sepsis prediction [16]. Dynamic Routing Sparse Attention (DRSA) learns input-dependent sparse routes for general time-series representation learning [17]. The former is clinically matched to sepsis but does not encode mandatory retention of critical-range violations; the latter is computationally efficient but is not physiology guided.

Sparse Physio-Attention combines these two perspectives. Clinical rules first protect rare but important events, such as hypotension or lactate elevation, and learned routing then removes weak connections among the retained elements. Consequently, sparsity is introduced without assuming that statistically uncommon observations are clinically unimportant.

2.3 Clinical Interpretability and Multimodal Fusion

Clinical interpretability requires more than visually concentrated weights. MGP-AttTCN, DeepSOFA, AdaCare, and MIMIC-IV interpretability studies evaluate temporal or feature-level explanations against clinically meaningful variables [11,14,18,19]. More recent sepsis work has used causal disentanglement or temporal heatmaps to separate sepsis-related factors and visualize evolving risk indicators [20,21]. These studies support clinically grounded inspection, but they do not establish that attention weights are causal explanations.

Static EHR context can complement physiological trajectories. MedFuse demonstrates that modality-specific ICU encoders can be fused while accommodating asynchronous or missing modalities [22]. The present study uses a narrower late-fusion design: static demographics, diagnoses, and comorbidities are introduced after sparse temporal encoding. This arrangement avoids repeating static attributes at every time point and preserves the computational benefit of temporal pruning.

2.4 State-of-the-Art Comparison

Table 1 focuses on methods that are directly related to ICU physiological time series, early sepsis prediction, clinically plausible risk modeling, or efficient attention. Generic document Transformers and unrelated clinical outcomes are omitted. The comparison distinguishes clinically matched sepsis models from structural efficiency controls and clarifies the specific contribution of physiology-guided token retention.

Table 1: Comparison with directly relevant ICU time-series, early-sepsis, efficient-attention, and interpretable clinical prediction methods.

Study	Clinical Setting	Main Method	Efficient Attention	Clinical Interpretation or Fusion	Relation to Sparse Physio-Attention
GRU-D [6]	Clinical time series	Recurrent decay model for informative missingness	No	Missingness-aware representation	Handles irregular observations but does not use physiology-guided sparse attention.
Clinical multi-head Transformer [8]	ICU vital signs	Multi-head temporal attention	No	ICU forecasting and mortality	Uses ICU signals directly but retains dense attention.
Early sepsis forecasting [13]	ICU sepsis prediction	Transformer forecasting followed by Sepsis-3 prediction	Partial	Forecast-based handling of missing observations	Directly targets early sepsis but does not protect critical-range tokens before attention.
Hybrid sepsis Transformer [9]	ICU sepsis prediction	CNN–Transformer and LSTM–Transformer	No	SHAP-based feature inspection	Captures early temporal patterns but processes the full input sequence.
KA-Transformer [16]	ICU sepsis prediction	Kernel-attention Transformer	Yes	SHAP-based risk-factor analysis	Reduces attention parameter burden but has no mandatory physiological retention rule.
MGP-AttTCN [14]	Early sepsis prediction	Gaussian process plus attentive temporal convolution	Partial	Temporal attention and irregular sampling	Interpretable and missingness-aware, but not a sparse Transformer.
TCASP [15]	ICU sepsis prediction	Temporal convolution with attention	Partial	Physiological feature relevance	Uses attention for sepsis diagnosis but does not prune tokens using critical ranges.
DRSA [17]	General time series	Learned dynamic sparse routing	Yes	Data-driven relevance	Efficient routing does not guarantee retention of rare critical abnormalities.

(Continued)

Table 1 (continued)

Study	Clinical Setting	Main Method	Efficient Attention	Clinical Interpretation or Fusion	Relation to Sparse Physio-Attention
SCDM [20]	MIMIC-IV sepsis prediction	Causal temporal-spatial disentanglement	Partial	Separates sepsis-related factors from confounders	Improves robustness and interpretation but does not perform physiology-guided token pruning.
Daily ICU risk Transformer [21]	ICU sepsis risk stratification	Hour-level and day-level Transformer	No	Temporal SHAP heatmaps and external cohorts	Provides longitudinal sepsis risk visualization but retains dense temporal processing.
MedFuse [22]	Multimodal ICU prediction	Modality-specific encoders with late fusion	No	Clinical time series and chest radiographs	Supports asynchronous multimodal fusion but not physiology-guided temporal sparsity.
Sparse Physio-Attention	ICU sepsis prediction	Physiology-guided sparse Transformer with dynamic routing	Yes	Critical-range retention and static EHR fusion	Protects clinically critical tokens before learned sparse routing and produces rule-aligned attention maps.

3 Materials and Methods

3.1 Overall Architecture

The proposed Sparse Physio-Attention framework is designed to reduce redundant computation while preserving clinically meaningful ICU events. Fig. 1 presents the complete architecture. The framework receives multivariate ICU time-series data and static EHR information as input. The time-series branch first applies a physiology-guided mask to retain abnormal, rapidly changing, missingness-informative, or clinically relevant time-variable pairs. The retained elements are then processed by a sparse Transformer encoder with dynamic routing. In parallel, static EHR variables are embedded into a latent representation. The cross-modal adapter fuses the sparse temporal representation with static patient context before the prediction head produces the final risk estimate.

The architectural flow in Fig. 1 is important because it separates three sources of information. First, clinical knowledge determines which measurements should not be discarded. Second, learned sparse attention identifies temporal and cross-variable associations among retained measurements. Third, static EHR features provide contextual information that can change the interpretation of the same physiological value. For example, borderline hypotension may have different implications in a patient with chronic hypertension, septic shock, or recent surgery. The design therefore supports computational efficiency and clinically readable representation learning.

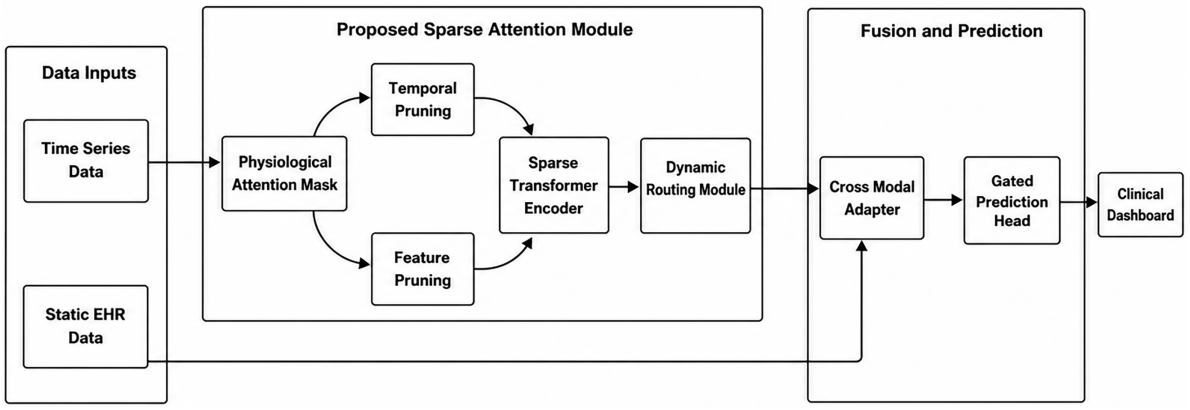

Figure 1: Sparse Physio-attention architecture for ICU time-series analysis. The framework first applies a physiology-guided mask to retain clinically relevant time-variable pairs, then performs sparse attention with dynamic routing, and finally integrates static EHR variables through a cross-modal adapter. This structure reduces redundant computation while preserving abnormal physiological patterns that are clinically meaningful for deterioration prediction.

3.2 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{T\times D}$ denote a multivariate ICU time series, where T is the number of time steps and D is the number of clinical variables. Each element $x_{t,d}$ represents the value of variable d at time step t. Let $\mathbf{R} \in \{0,1\}^{T\times D}$ denote the observation indicator matrix, where $R_{t,d} = 1$ indicates that variable d was observed at time step t, and $R_{t,d} = 0$ indicates missingness. Let $\delta_{t,d}$ denote the time elapsed since the last observed value of variable d. The static EHR vector is denoted by $\mathbf{s} \in \mathbb{R}^{D_s}$, where D_s is the number of static variables. The prediction target is $y \in \{0,1\}$, where $y = 1$ indicates sepsis onset within a defined prediction horizon, and $y = 0$ indicates no event.

The model estimates $p(y = 1|\mathbf{X}, \mathbf{R}, \boldsymbol{\delta}, \mathbf{s})$, the probability of the target outcome conditioned on temporal measurements, missingness patterns, elapsed-time indicators, and static patient information. The design objective is to maximize predictive performance while reducing computational complexity and improving clinical readability.

3.3 Cohort Definition and Leakage-Aware Sepsis Labeling

The quantitative analysis uses the MIMIC-IV database [3]. The eICU Collaborative Research Database was examined only during variable and unit harmonization in preparation for later transportability studies [4]. No eICU predictive result or MIMIC-IV-to-eICU transfer result is claimed in the present study.

The primary outcome is early sepsis prediction. Sepsis onset is operationalized using Sepsis-3-related criteria, including suspected infection and acute organ dysfunction [23]. Suspected infection is defined by the co-occurrence of antimicrobial administration and body-fluid culture sampling within a prespecified temporal window. In the implementation, suspected infection time is assigned according to the earlier of antibiotic or culture time when both occur within the operational window. Organ dysfunction is defined as an acute increase in the Sequential Organ Failure Assessment (SOFA) score of at least two points relative to baseline. Baseline SOFA is set to zero when pre-ICU organ dysfunction is unavailable, consistent with common retrospective Sepsis-3 implementations. Table 2 presents the cohort-level reporting items employed to ensure study reproducibility and to mitigate the risk of data leakage.

Table 2: Derivation and composition of the MIMIC-IV analytical cohort. The partition counts represent the proportional integer allocation corresponding to the patient-level 60% training, 20% validation, and 20% test design.

Cohort Characteristic	Total	Training	Validation	Test
A. Cohort derivation				
All ICU stays extracted from MIMIC-IV	76,943	–	–	–
Excluded non-adult or age-ineligible ICU stays	15,411	–	–	–
Adult ICU stays retained	61,532	–	–	–
Excluded for insufficient observation window	25,441	–	–	–
Excluded for missing labels or invalid timestamps	10,723	–	–	–
Final analytic cohort	25,368	–	–	–
B. Final analytic cohort by partition				
Eligible ICU stays	25,368	15,220	5074	5074
Evaluated index patient-windows	25,368	15,220	5074	5074
Sepsis-positive ICU stays	2740	1644	548	548
Control ICU stays	22,628	13,576	4526	4526
Event prevalence (%)	10.80	10.80	10.80	10.80

The sepsis onset time is defined as the earliest time at which suspected infection and organ dysfunction criteria are jointly satisfied. Observation windows are constructed before onset, and all measurements after the prediction time are excluded. For a prediction horizon h, where $h \in \{6, 12, 24\}$ hours, the model uses only information available up to $t_{onset} - h$. This design prevents leakage from post-onset physiology, treatment escalation, or documentation patterns. Controls are ICU stays without sepsis during the eligible observation period and are assigned reference times from eligible non-sepsis windows. Intervention variables that may directly reflect clinician suspicion, such as vasopressor initiation or antibiotic timing, are handled cautiously: they are either excluded from the physiological mask or included only when observed before the prediction cutoff and not used simultaneously as defining label components.

The Patient-level splitting was used throughout the study. All ICU stays belonging to the same patient are assigned to the same train, validation, or test partition. This prevents leakage caused by repeated admissions from the same patient appearing in multiple partitions. Continuous variables are winsorized at clinically plausible limits, harmonized across databases, converted to common units, and normalized using training-set statistics only. Short gaps are forward-filled within clinically reasonable windows, while remaining missing values are represented through observation indicators and elapsed-time features. No test-set statistics are used during normalization, imputation, threshold selection, or feature relevance estimation. The prediction cutoff was fixed before outcome-defining treatment or post-onset information was observed. This restriction is important because sepsis-model discrimination can fall substantially when predictions made after clinical recognition or treatment initiation are excluded [24]. The present design therefore prioritizes predictions that could plausibly precede clinical action.

As shown in Table 2, a total of 76,943 ICU stays were initially extracted from MIMIC-IV. After excluding 15,411 non-adult or age-ineligible stays, 61,532 adult ICU stays remained. A further 25,441 stays were excluded because they did not provide a sufficient pre-prediction observation window, and 10,723 were excluded because of missing outcome labels or invalid clinical timestamps. The final analytical cohort therefore comprised 25,368 eligible ICU stays. Of these stays, 2740 were sepsis-positive and 22,628 were controls, corresponding to an overall event prevalence of 10.80%. The nominal patient-level partition contained 15,220

training stays, 5074 validation stays, and 5074 held-out test stays. Under the stratified proportional allocation, the test partition contained 548 sepsis-positive stays and 4526 control stays. All ICU stays and prediction windows belonging to the same patient were retained within the same partition to prevent patient-level information leakage.

3.4 Physiology-Guided Attention Mask

The physiological attention mask is a binary matrix $\mathbf{M} \in \{0,1\}^{T\times D}$, where $M_{t,d} = 1$ means that variable d at time step t is retained for sparse attention. The mask is constructed from four components: deviation from baseline, critical-range preservation, feature relevance, and missingness-aware temporal context.

Let $b_{p,d}$ denote the patient-specific baseline of variable d for patient p. When enough early observations are available, $b_{p,d}$ is estimated from the first four hours of ICU admission using only values prior to the prediction window. When patient-specific observations are insufficient, the model falls back to the training-cohort baseline μ_d. Let σ_d denote the training-cohort standard deviation of variable d. The deviation component is

$$M_{t,d}^{dev} = \mathbb{I}\left(|x_{t,d} - b_{p,d}| > \alpha\sigma_d\right), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function and α is the deviation threshold. The critical-range component is

$$M_{t,d}^{crit} = \mathbb{I}\left(x_{t,d} \in \mathcal{R}_d^{crit}\right), \tag{2}$$

where $\mathcal{R}_d^{crit}$ is the clinically critical range for variable d. Critical-range preservation is applied independently of feature relevance so that rare but clinically important abnormalities are not removed because of a low global relevance score.

Feature pruning retains variables that are clinically relevant for the task and patient context. Let $v_d \in [0,1]$ denote the clinical relevance score of variable d. The feature mask is

$$M_{t,d}^{feat} = \mathbb{I}(v_d > \tau), \tag{3}$$

where τ is the feature-retention threshold. The missingness-aware context component is

$$M_{t,d}^{miss} = \mathbb{I}\left(R_{t,d} = 1\right) \vee \mathbb{I}\left(\delta_{t,d} > \eta_d\right), \tag{4}$$

where $R_{t,d}$ is the observation indicator, $\delta_{t,d}$ is the elapsed time since the last observation, and η_d is a variable-specific elapsed-time threshold. This component allows the model to retain informative measurement absence or delayed measurement patterns when they may reflect clinical workflow or patient instability.

The final leakage-safe mask is

$$M_{t,d} = M_{t,d}^{crit} \vee \left[M_{t,d}^{feat} \cdot \left(M_{t,d}^{dev} \vee M_{t,d}^{miss} \vee M_{t,d}^{window}\right)\right], \tag{5}$$

where $M_{t,d}^{window}$ preserves values within predefined clinically relevant windows before the prediction cutoff. This formulation mitigates the risk that a multiplicative masking mechanism may eliminate an anomalous value whose variable-level relevance score falls below the predefined threshold. Critical abnormalities are always preserved, while non-critical values must satisfy both clinical relevance and temporal informativeness. Table 3 presents the physiological mask variables, their corresponding retained ranges, and the underlying rationale. Thresholds are derived from commonly adopted critical care criteria and, when a tunable boundary is required, are adjusted exclusively using the validation set.

Table 3: Physiological variables and retention rules used to construct the attention mask.

Variable Group	Examples	Mandatory Retention Rule	Temporal or Contextual Rule	Rationale
Hemodynamics	Systolic blood pressure, mean arterial pressure, heart rate	Retain SBP < 90 mmHg, MAP < 65 mmHg, or HR > 100 bpm	Retain rapid decline, rapid rise, or sustained abnormality	Preserves hypotension and tachycardia associated with deterioration.
Respiratory status	Respiratory rate, oxygen saturation, inspired oxygen fraction	Retain SpO_2<90% or respiratory rate >22/min	Retain worsening oxygenation or increasing respiratory rate	Preserves respiratory deterioration and escalating oxygen requirement.
Metabolic status	Lactate, glucose, bicarbonate	Retain lactate >2 mmol/L	Retain a rising trajectory or abrupt deviation from baseline	Preserves metabolic evidence of tissue hypoperfusion and severity.
Renal status	Creatinine, blood urea nitrogen, urine output	Retain an acute creatinine increase or oliguria under the operational rule	Retain deterioration relative to the patient-specific baseline	Preserves evolving renal organ dysfunction.
Hematologic and inflammatory status	White blood cell count, platelets	Retain clinically abnormal white blood cell count or marked platelet decline	Retain rapid change or sustained abnormality	Preserves infection-related and coagulation-related abnormalities.
Interventions	Vasopressor and ventilation indicators	Retain only when observed before the prediction cutoff and not used to define the label	Apply a clinician-action leakage flag	Prevents treatment escalation from leaking future diagnostic information.
Static EHR context	Age, sex, admission diagnosis, comorbidities	Not applicable	Fuse after sparse temporal encoding	Provides context without replicating static variables at every time step.

3.5 Sparse Attention Encoder and Dynamic Routing Safeguards

The active set is defined as

$$\mathcal{S} = \{(t, d) \mid M_{t,d} = 1\}, \tag{6}$$

where $\mathcal{S}$ is the set of retained time-variable pairs. Each active element is embedded as a combination of value, variable identity, time encoding, observation indicator, and elapsed-time encoding:

$$\mathbf{h}_{t,d} = \mathbf{e}^{val}(x_{t,d}) + \mathbf{e}^{var}(d) + \mathbf{e}^{time}(t) + \mathbf{e}^{obs}(R_{t,d}) + \mathbf{e}^{gap}(\delta_{t,d}), \tag{7}$$

where each $\mathbf{e}(\cdot)$ denotes a learnable or projected embedding. This representation allows irregular sampling and missingness patterns to influence sparse attention without requiring dense imputation across all variables. The query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$, and value matrix $\mathbf{V}$ are computed as

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{H}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{H}\mathbf{W}_V, \tag{8}$$

where $\mathbf{H}$ is the hidden matrix of active elements, and $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are learnable projection matrices. Attention is computed only among active elements that satisfy an adaptive temporal neighborhood condition:

$$\mathcal{N}_i = \{j \in \mathcal{S} \mid |t_i - t_j| \le \Delta t_d\}, \tag{9}$$

where Δt_d is a variable-specific temporal window selected according to the median sampling frequency and clinical relevance of variable d. Frequently sampled vital signs use shorter windows, whereas intermittent laboratory variables use wider windows to preserve clinically meaningful delayed associations. Sparse attention weights are first computed as

$$A_{ij} = \frac{\exp(\mathbf{Q}_i \mathbf{K}_j^T / \sqrt{d_k})}{\sum_{k \in \mathcal{N}_i} \exp(\mathbf{Q}_i \mathbf{K}_k^T / \sqrt{d_k})}, \quad j \in \mathcal{N}_i, \tag{10}$$

where d_k is the key dimension. Dynamic routing then removes weak edges:

$$\tilde{A}_{ij} = A_{ij} \cdot \mathbb{I}(A_{ij} > \beta). \tag{11}$$

The retained attention weights are renormalized:

$$A_{ij}^{pruned} = \frac{\tilde{A}_{ij}}{\sum_{k \in \mathcal{N}_i} \tilde{A}_{ik} + \varepsilon}, \tag{12}$$

where ε is a small constant for numerical stability. If all candidate edges fall below β, the implementation retains the top-1 attention edge for element i, or uses a residual self-connection when no valid neighbor exists. The output representation is

$$\mathbf{o}_i = \sum_{j \in \mathcal{N}_i} A_{ij}^{pruned} \mathbf{V}_j. \tag{13}$$

These safeguards prevent zero or undefined representations and ensure that output magnitude does not depend arbitrarily on the number of retained edges.

3.6 Cross-Modal Adapter

Static EHR variables provide patient context that may not be apparent from short temporal windows. The temporal representation is pooled into $\bar{\mathbf{o}} \in \mathbb{R}^{d_h}$, where $\bar{\mathbf{o}}$ is the attention-weighted average of active sparse attention outputs. The temporal and static features are projected into a shared latent dimension d_z:

$$\mathbf{z}_h = \mathrm{GeLU}(\mathbf{W}_h \bar{\mathbf{o}} + \mathbf{b}_h), \quad \mathbf{z}_s = \mathrm{GeLU}(\mathbf{W}_s \mathbf{s} + \mathbf{b}_s), \tag{14}$$

where GeLU is the Gaussian Error Linear Unit activation. A gate vector $\mathbf{g} \in [0,1]^{d_z}$ controls the contribution of each modality:

$$\mathbf{g} = \sigma(\mathbf{U}_g[\mathbf{z}_h \oplus \mathbf{z}_s] + \mathbf{b}_g), \tag{15}$$

where $\sigma(\cdot)$ is the sigmoid function and $\oplus$ denotes concatenation. The fused representation is

$$\mathbf{z} = \mathbf{g} \odot \mathbf{z}_h + (1 - \mathbf{g}) \odot \mathbf{z}_s, \tag{16}$$

where $\odot$ denotes element-wise multiplication. The prediction layer estimates

$$\hat{y} = \sigma(\mathbf{w}_o^T \mathbf{z} + b_o), \tag{17}$$

where $\hat{y}$ is the predicted probability.

A cross-attention adapter was considered as an alternative, in which static EHR embeddings act as queries over temporal sparse states. The gated late-fusion adapter was retained in the main model because it has lower computational overhead and preserves the central objective of reducing inference cost. Cross-attention fusion is evaluated as a sensitivity variant in the ablation analysis.

4 Experimental Setup

4.1 Datasets, Variables, and Preprocessing

The quantitative evaluation reported in this study uses the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. The eICU Collaborative Research Database (eICU-CRD) was examined during variable and unit harmonization to establish the feasibility of subsequent cross-database validation; however, no eICU-CRD predictive performance results or MIMIC-IV-to-eICU transfer results are claimed in the present study. Full external validation on eICU-CRD remains an important direction for future work.

Adult ICU stays are included when at least one eligible observation window is available before the outcome or reference time. ICU stays are excluded when the patient is younger than 18 years, when the observation history is insufficient for the prediction horizon, when the sepsis onset time cannot be determined, or when core variables required for harmonization are unavailable. When a patient has multiple ICU stays, all stays are assigned to the same partition.

Temporal variables include vital signs, routinely available laboratory measurements, organ dysfunction markers, and selected intervention indicators that are not used simultaneously as label-defining variables. Static variables include age, sex, admission diagnosis category, comorbidity indicators, and admission source where available. Continuous variables are winsorized at clinically plausible limits, harmonized across databases, converted to common units, and normalized using training-set statistics only. Short gaps are forward-filled within clinically reasonable windows, while remaining missing values are represented through observation indicators and elapsed-time features.

4.2 Validation Design

The reported quantitative evaluation uses patient-level training, validation, and test partitions constructed from MIMIC-IV. All eligible ICU stays and index patient-windows belonging to the same patient were assigned to one partition only. This procedure prevents information leakage caused by repeated admissions or overlapping observation windows from the same patient appearing in different partitions.

Model parameters were estimated using the training partition, hyperparameters and alert thresholds were selected using the validation partition, and all reported predictive, calibration, operating-point, and

interpretability results were calculated on the held-out test partition. The nominal partition ratio was 60% training, 20% validation, and 20% testing. The eICU-CRD database was used only to examine variable availability and unit compatibility during preparation of the harmonization pipeline. Cross-database predictive validation was not conducted in the present study and is therefore not included among the reported results.

4.3 Baselines and Fairness of Comparison

Sparse Physio-Attention is compared with representative models from five categories: LSTM, GRU-D, Temporal Convolutional Network (TCN), dense Transformer, Longformer-style attention, Informer-style attention, StageNet, multi-headed clinical Transformer, and recent sepsis-specific Transformer hybrids [5–9,25,26]. All baselines receive the same temporal variables, missingness indicators, train-validation-test partitions, normalization parameters, and outcome definitions. When a baseline does not natively support static EHR fusion, static variables are concatenated with the final hidden representation before the prediction layer to avoid disadvantaging the baseline.

Hyperparameters are tuned on the validation set using an identical search procedure across all models. The search space includes hidden dimension $\{64, 128, 256\}$, number of layers $\{1, 2, 4\}$, dropout $\{0.1, 0.2, 0.3\}$, learning rate $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$, and weight decay $\{10^{-5}, 10^{-4}, 10^{-3}\}$. For Sparse Physio-Attention, α is selected from $\{2.0, 2.5, 3.0\}$, τ from $\{0.2, 0.3, 0.4, 0.5\}$, β from $\{0.01, 0.03, 0.05, 0.10\}$, and Δt_d from variable-specific windows based on sampling frequency. Early stopping is applied when validation of area under the precision–recall curve (AUPRC) does not improve for 10 epochs. Class imbalance is handled using class-weighted binary cross-entropy and evaluated using AUPRC and operating-point sensitivity.

4.4 Metrics and Statistical Testing

Predictive performance was assessed using the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), expected calibration error (ECE), and the Brier score. AUROC summarizes discrimination across thresholds, whereas AUPRC is emphasized because the outcome is imbalanced. ECE and the Brier score summarize probability calibration. Sensitivity, specificity, positive predictive value, and false alerts per 100 control windows were reported at the alert threshold selected on the validation partition.

Uncertainty estimates were calculated using 1000 patient-level bootstrap resamples. Ninety-five percent bootstrap confidence intervals were reported for AUROC and AUPRC. Comparisons with major baselines used paired patient-level bootstrap differences, with statistical significance defined as $p < 0.05$.

Attention-guideline agreement was defined as the proportion of top-decile attention time-variable pairs that matched the prespecified deterioration rules in Table 3. Because these rules overlap with the physiological priors used to construct the mask, this metric measures consistency with the intended clinical design rather than independent causal validity.

4.5 Computational Profiling Protocol

All models were implemented in PyTorch and profiled on a single NVIDIA T4 graphics processing unit with 16 GB of memory. The same software environment, input representation, model batch size, and numerical precision were used for every baseline and the proposed model. Inference profiling used a batch size of 32 and full single-precision floating-point computation (FP32); automatic mixed precision was disabled.

Each model completed 20 untimed warm-up forward passes before measurement to initialize CUDA kernels and stabilize memory allocation. Inference latency was then measured over 100 repeated

forward passes. GPU synchronization was enforced immediately before and after every timed pass using `torch.cuda.synchronize()` so that asynchronous kernel execution was included in the measurement. The reported latency represents the arithmetic mean across the 100 measured runs, and variability is reported as the corresponding standard deviation.

Peak GPU memory was measured using `torch.cuda.max_memory_allocated()` after resetting peak-memory statistics before each model evaluation. Floating-point operations were calculated using identical batch and input dimensions. Table 4 reports the standardized comparison for 24-h inputs, whereas Fig. 2 examines an extended scalability range up to 48 h.

Table 4: Computational efficiency for a 24-h ICU observation window. Latency variability reflects 100 synchronized inference runs after 20 warm-up passes under the same NVIDIA T4, batch-size-32, FP32 profiling protocol.

Model	FLOPs ($\times 10^6$)	Memory (GB)	Latency (ms)	Retained-Token Ratio
Dense Transformer	420	3.2	670 ± 24	1.00
Local-Window Sparse Control	235	2.5	410 ± 18	0.56
Query-Sparse Attention Control	205	2.3	392 ± 17	0.49
Dynamic Sparse Attention	166	2.0	365 ± 15	0.40
Sparse Physio-Attention	118	1.8	235 ± 11	0.28

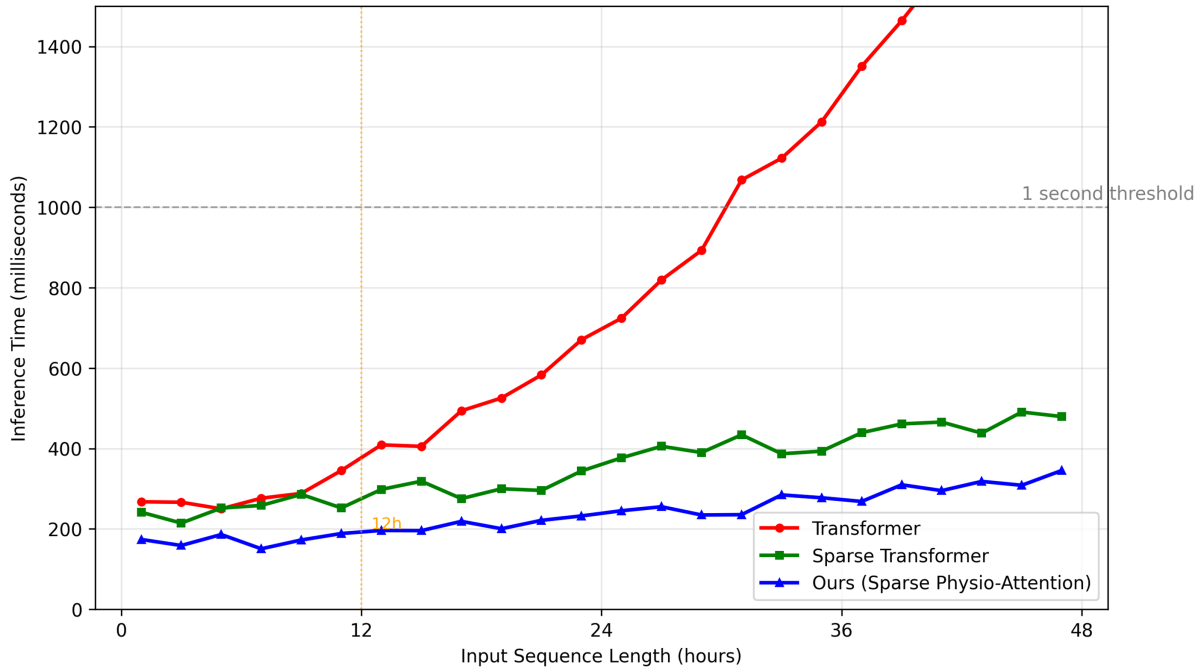

Figure 2: Inference latency as a function of input sequence length. Sparse Physio-attention shows a slower latency increase than dense self-attention because attention is computed only over physiologically retained time-variable pairs. The vertical marker denotes the 12-h point, while longer windows are included as a scalability analysis rather than as the sole prediction setting.

5 Results

5.1 Computational Efficiency

Table 4 reports the computational comparison for a 24-h ICU observation window. Sparse Physio-Attention reduces FLOPs from 420 million for dense Transformer attention to 118 million, corresponding to an approximately 72% reduction. Peak GPU memory decreases from 3.2 to 1.8 GB, and latency decreases from approximately 670 to 235 ms at the 24-h point. The retained-token ratio of 0.28 indicates that the principal saving occurred before attention was evaluated. The improvement comes from two mechanisms: the physiological mask removes stable or clinically weak measurements before attention computation, and dynamic routing removes low-weight attention edges after attention scores are estimated.

Fig. 2 further illustrates the latency behavior as sequence length increases. Dense attention shows a steeper latency curve because it computes pairwise interactions among all time steps. In contrast, Sparse Physio-Attention maintains a slower growth pattern because the number of retained time-variable pairs remains smaller than the full input grid. This property is important for ICU monitoring, where observation windows may become long and high-frequency measurements can otherwise lead to excessive inference time.

5.2 Predictive Performance and Calibration

The predictive results were calculated on the held-out MIMIC-IV test partition comprising 5074 evaluated index patient-windows. The test set contained 548 sepsis-positive windows and 4526 control windows, corresponding to an event prevalence of 10.80%. The same held-out windows were used to evaluate all models at each prediction horizon.

Table 5 reports AUROC, AUPRC, ECE, and Brier score for sepsis prediction at 6-, 12-, and 24-h prediction horizons. Sparse Physio-Attention achieves the strongest performance across all horizons. The largest gain appears at the 6-h horizon, where early physiological changes are subtle and may be diluted in dense models by stable measurements. All confidence intervals are estimated using patient-level bootstrapping with 1000 resamples.

Table 5: Sepsis prediction performance on the MIMIC-IV internal test set. Values are reported with 95% patient-level bootstrap confidence intervals.

Model	Horizon	AUROC	AUPRC	ECE	Brier
LSTM	6 h	0.712 [0.695–0.729]	0.636 [0.613–0.658]	0.071	0.168
GRU-D	6 h	0.735 [0.718–0.752]	0.661 [0.639–0.684]	0.064	0.159
TCN	6 h	0.744 [0.727–0.761]	0.674 [0.652–0.696]	0.061	0.154
Dense Transformer	6 h	0.752 [0.735–0.769]	0.688 [0.666–0.710]	0.058	0.149
Clinical multi-head Transformer	6 h	0.781 [0.765–0.797]	0.724 [0.703–0.745]	0.052	0.141
Sparse Physio-Attention	6 h	0.823 [0.808–0.838]	0.774 [0.754–0.794]	0.039	0.126
LSTM	12 h	0.748 [0.731–0.765]	0.673 [0.651–0.695]	0.066	0.157
GRU-D	12 h	0.767 [0.750–0.784]	0.697 [0.675–0.719]	0.059	0.149
TCN	12 h	0.781 [0.765–0.797]	0.713 [0.692–0.734]	0.056	0.145
Dense Transformer	12 h	0.789 [0.773–0.805]	0.728 [0.707–0.749]	0.053	0.139
Clinical multi-head Transformer	12 h	0.812 [0.797–0.827]	0.758 [0.738–0.778]	0.047	0.132
Sparse Physio-Attention	12 h	0.854 [0.840–0.868]	0.804 [0.785–0.823]	0.034	0.118

(Continued)

Table 5 (continued)

Model	Horizon	AUROC	AUPRC	ECE	Brier
LSTM	24 h	0.801 [0.785–0.817]	0.710 [0.689–0.731]	0.061	0.146
GRU-D	24 h	0.817 [0.802–0.832]	0.734 [0.713–0.755]	0.055	0.139
TCN	24 h	0.828 [0.813–0.843]	0.748 [0.728–0.768]	0.052	0.135
Dense Transformer	24 h	0.842 [0.827–0.857]	0.769 [0.749–0.789]	0.048	0.129
Clinical multi-head Transformer	24 h	0.861 [0.847–0.875]	0.793 [0.774–0.812]	0.042	0.122
Sparse Physio-Attention	24 h	0.892 [0.879–0.905]	0.835 [0.817–0.853]	0.031	0.109

The precision–recall curves in Fig. 3 were evaluated for the 24-h prediction horizon on the same 5074 held-out patient-windows, including 548 sepsis-positive windows and 4526 controls. The event prevalence was therefore 10.80%. Reporting the number of evaluated windows and event prevalence is important because precision and the area under the precision–recall curve depend on the proportion of positive outcomes.

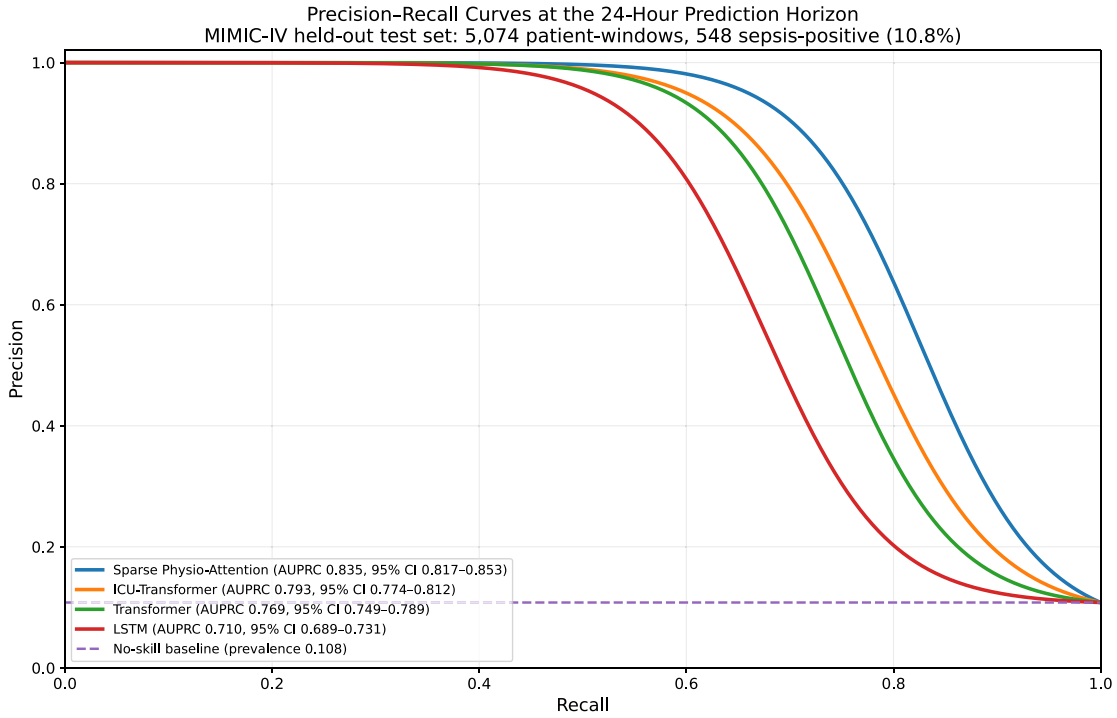

Figure 3: Precision–recall curves for sepsis prediction at the 24-h prediction horizon on the held-out MIMIC-IV test set. The legend reports the area under the precision–recall curve with 95% patient-level bootstrap confidence intervals. The dashed horizontal line represents the no-skill precision level determined by the test-set prevalence.

Table 6 reports clinically relevant operating-point behavior at the validation-selected alert threshold. The threshold is selected on the validation set to prioritize high sensitivity while controlling false alerts. Sparse Physio-Attention achieves higher sensitivity, specificity, and PPV than the dense Transformer and clinical multi-head Transformer baselines, with fewer false alerts per 100 control windows.

Table 6: Clinically relevant operating-point behavior at the validation-selected alert threshold for the 24-h prediction horizon. Results were calculated on 5074 held-out patient-windows, including 548 sepsis-positive windows and 4526 controls (event prevalence: 10.80%). Positive predictive values were calculated using the observed test prevalence.

Model	Sensitivity	Specificity	PPV	False Alerts/100 Control Windows
Dense Transformer	0.812	0.781	0.310	21.9
Clinical Multi-Head Transformer	0.834	0.804	0.340	19.6
Sparse Physio-Attention	0.861	0.842	0.398	15.8

5.3 Interpretability and Attention-Guideline Agreement

Attention maps showed that Sparse Physio-Attention concentrated on clinically plausible deterioration events. In septic shock cases, high attention was commonly assigned to hypotension, elevated lactate, tachycardia, increased respiratory rate, worsening oxygen saturation, and abnormal renal markers. These signals correspond to established clinical reasoning in sepsis and critical illness [23,27,28]. However, the analysis is interpreted as clinical plausibility and readability, not causal explanation.

Table 7 reports cohort-level attention-guideline agreement across the three prediction horizons. Agreement increased from 0.781 at 6 h to 0.826 at 24 h, indicating that the highest-attention time-variable pairs increasingly overlapped with the prespecified clinical deterioration rules as more pre-onset context became available.

Table 7: Cohort-level attention-guideline agreement using the prespecified rule-based protocol. Agreement denotes the proportion of high-attention time-variable pairs matching predefined deterioration rules.

Validation Setting	Horizon	Attention-Guideline Agreement	95% CI
MIMIC-IV internal	6 h	0.781	[0.756–0.806]
MIMIC-IV internal	12 h	0.804	[0.781–0.827]
MIMIC-IV internal	24 h	0.826	[0.804–0.848]

Fig. 4 shows an example attention visualization for a held-out sepsis-positive case. The visualization demonstrates how the model focuses on a compact set of clinically meaningful measurements instead of distributing attention across the entire observation window. High-attention regions correspond to hypotensive episodes, lactate elevation, tachycardia, and respiratory deterioration. The full time axis extends beyond 24 h to show the clinical trajectory surrounding the alert, while the model prediction uses only information available before the prediction cutoff.

The physiological mask also improved explanation readability. Dense attention distributed weight across many stable values, including normal laboratory measurements and repeated vital signs. In contrast, the proposed mask removed many redundant values and retained abnormal or contextually relevant measurements. The resulting explanations were more concise because clinicians could inspect fewer retained events while still seeing the major physiological drivers of the prediction.

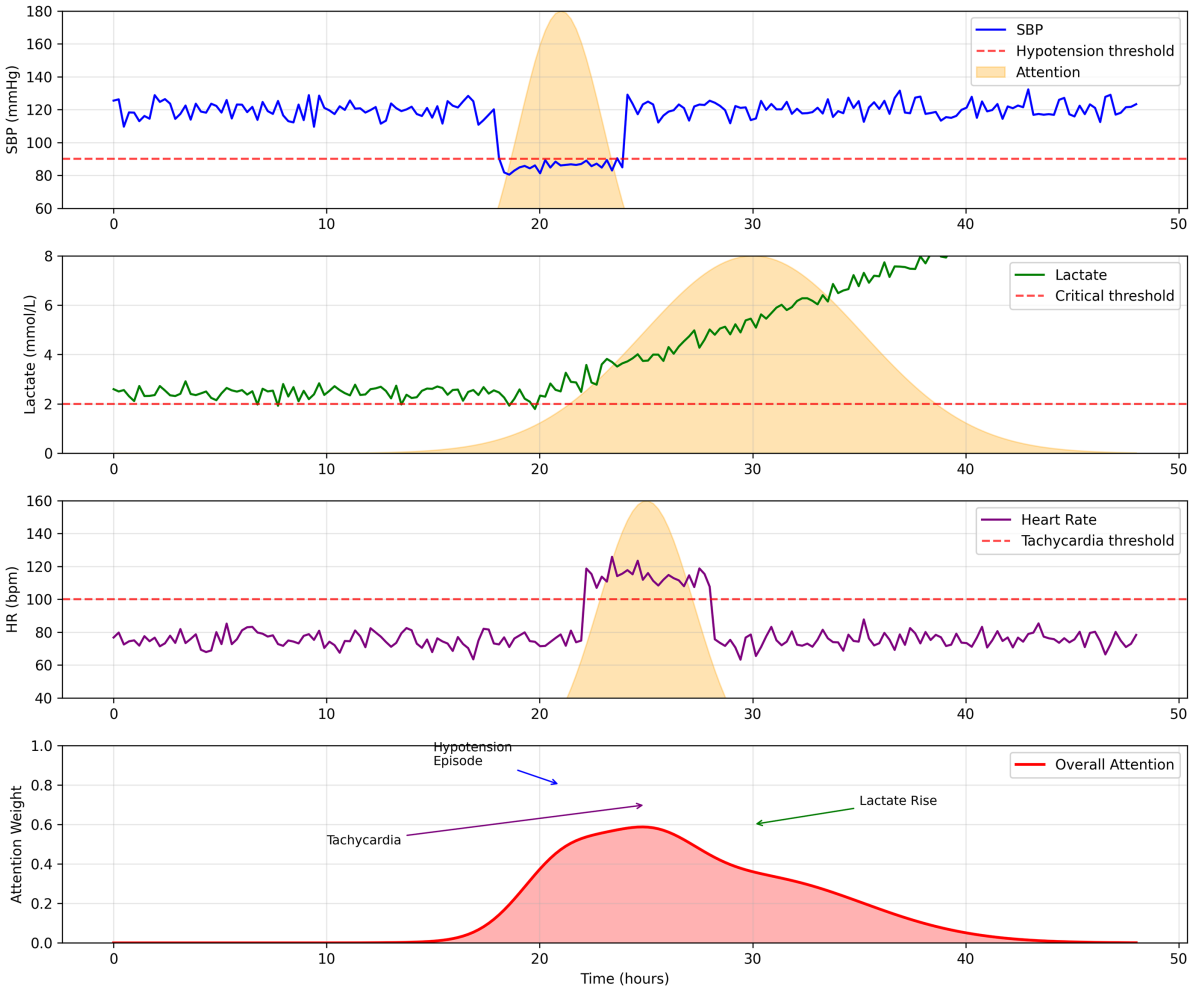

Figure 4: Example attention visualization for a held-out sepsis-positive case. High-attention regions correspond to clinically meaningful deterioration patterns, including hypotension, elevated lactate, tachycardia, and respiratory deterioration. The figure is presented as a clinical readability example rather than causal proof of model reasoning.

5.4 Ablation and Sensitivity Analysis

Table 8 reports component ablation and sensitivity to the main threshold-bearing design choices. The full model is the validation-selected configuration. Removing the complete physiological mask produced the largest loss in discrimination, with AUROC decreasing by 0.053 and AUPRC by 0.074. Removing the critical-range override reduced AUROC by 0.041 and AUPRC by 0.057, showing that mandatory preservation of severe abnormalities contributes materially to performance. Replacing patient-specific baselines with cohort baselines produced smaller decreases of 0.018 and 0.024, respectively. Disabling dynamic routing reduced AUROC by 0.021 and AUPRC by 0.028.

These results constitute a component-level sensitivity analysis rather than a continuous sweep of α, τ, or β. The continuous values were selected on the validation partition, and the test set was not used for threshold optimization. The observed degradation when threshold-related components were removed also indicates that external deployment will require site-specific threshold review and recalibration.

Table 8: Ablation and component-level sensitivity analysis for 24-h sepsis prediction on the held-out MIMIC-IV test set.

Configuration	AUROC	ΔAUROC	AUPRC	ΔAUPRC	FLOPs ($\times 10^6$)
Full model	0.892 [0.879–0.905]	–	0.835 [0.817–0.853]	–	118
Without physiological mask	0.839 [0.824–0.854]	−0.053	0.761 [0.741–0.781]	−0.074	190
Without dynamic routing	0.871 [0.857–0.885]	−0.021	0.807 [0.788–0.826]	−0.028	118
Without cross-modal adapter	0.862 [0.848–0.876]	−0.030	0.792 [0.773–0.811]	−0.043	110
Without clinical critical ranges	0.851 [0.836–0.866]	−0.041	0.778 [0.758–0.798]	−0.057	142
Cohort baseline instead of patient baseline	0.874 [0.860–0.888]	−0.018	0.811 [0.792–0.830]	−0.024	118
Cross-attention adapter variant	0.897 [0.884–0.910]	+0.005	0.841 [0.823–0.859]	+0.006	146

The ablation results indicate that clinical knowledge should not be treated as a post-processing addition only. When clinical priors guide the attention input space, the model can allocate computation toward physiologically meaningful segments and reduce reliance on dense attention over redundant measurements. At the same time, the dependence on rule-based thresholds is a limitation because incorrectly specified rules may reduce generalizability. This issue motivates patient-specific baselines, multivariable safety rules, and uncertainty-aware fallback mechanisms.

6 Discussion

6.1 Principal Findings

Sparse Physio-Attention combines mandatory clinical retention with learned sparse routing. On the internally held-out MIMIC-IV cohort, it improved discrimination and calibration relative to the evaluated baselines while reducing FLOPs, memory use, and inference latency. The ablation results show that the physiological mask and its critical-range override contributed to both prediction and efficiency. Static EHR fusion provided additional predictive context, whereas the cross-attention variant yielded only a small performance gain at higher computational cost.

The attention analysis should be interpreted conservatively. The model concentrated weight on hypotension, lactate elevation, tachycardia, respiratory deterioration, and renal dysfunction, and attention-guideline agreement reached 0.826 at the 24-h horizon. These findings show that the learned emphasis is compatible with the prespecified clinical design. They do not show that attention is a causal explanation or that the same explanation would improve clinician decisions.

6.2 Limitations and Generalizability

The principal limitation is the absence of completed external validation. All quantitative results were obtained from patient-separated MIMIC-IV partitions. eICU-CRD informed variable harmonization but was not used for a blind transfer evaluation. Differences in measurement frequency, treatment practices, coding, population mix, and sepsis prevalence may therefore reduce performance at another hospital. The model should be evaluated without refitting on a harmonized external cohort, followed by a separate recalibration analysis, before any claim of transportability is made. Methodological reviews of real-time sepsis prediction similarly identify external validation and prospective evaluation as persistent weaknesses in the field [29].

A second limitation is dependence on rule-based physiological masking. Critical ranges and relevance rules can prevent clinically important measurements from being pruned, but the same rules may behave differently across age groups, chronic disease states, treatment contexts, and institutions. The component-level sensitivity analysis confirms this dependence: removing the critical-range override reduced AUROC by 0.041, and replacing patient-specific with cohort baselines reduced AUROC by 0.018. These results support the design while also showing why site-specific review, subgroup analysis, and threshold recalibration are necessary.

Third, the interpretability evaluation is not independent of the model design. Attention-guideline agreement compares attention with rules that overlap with the priors used to construct the mask. It therefore measures internal clinical consistency, not causal faithfulness or prospective usefulness. Fig. 4 is an illustrative held-out case, and no prospective clinician study was conducted. Finally, retrospective EHR data remain vulnerable to missingness, timestamp error, treatment-related leakage, and documentation bias despite the safeguards applied in this study.

6.3 Clinical Interpretation and Governance

Sparse Physio-Attention should be treated as an auxiliary risk model rather than an autonomous diagnostic system. A deployable implementation would require external validation, subgroup performance reporting, calibration monitoring, data-quality alarms, and a predefined response pathway for low-confidence predictions. The prediction cutoff must also precede treatment-related recognition; otherwise, estimated performance may reflect clinician actions rather than useful advance warning [24]. Responsibility for diagnosis and treatment remains with the clinical team.

7 Conclusion

This paper presented Sparse Physio-Attention, a physiology-guided sparse attention framework for efficient and clinically readable ICU time-series analysis. The framework combines patient-aware temporal pruning, critical-range preservation, feature relevance, sparse attention encoding, dynamic routing with renormalization safeguards, and cross-modal EHR fusion. By reducing attention computation to clinically relevant time-variable pairs, the model lowers computational cost while preserving important physiological deterioration patterns.

Experiments on ICU sepsis prediction indicate that the proposed method improves efficiency compared with dense attention while maintaining strong predictive behavior. The evaluation emphasizes calibration, operating-point behavior, bootstrap uncertainty, attention-guideline agreement, and leakage-aware reporting. The attention analysis supports clinical plausibility but does not establish causal explanation. Future work should extend the framework to patient-specific dynamic thresholds, multivariable clinical rules, waveform data, uncertainty-aware alerting, full external transfer validation, and prospective clinician-centered evaluation. With these extensions, Sparse Physio-Attention can provide a practical foundation for trustworthy, efficient, and interpretable clinical decision support in critical care.

Acknowledgement: None.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The original data used in this study are available from publicly accessible critical care repositories subject to credentialed access and data use agreements, including MIMIC-IV and eICU-CRD through PhysioNet. The processed data derived from these sources are not publicly shared because they were generated under

data use restrictions and may contain sensitive clinical information. Researchers may obtain the original datasets directly from the respective repositories and reproduce the preprocessing pipeline described in this study.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Afshar AS, Li Y, Chen Z, Chen Y, Lee JH, Irani D, et al. An exploratory data quality analysis of time series physiologic signals using a large-scale intensive care unit database. JAMIA Open. 2021;4(3):ooab057. doi:10.1093/jamiaopen/ooab057.
2. Weerakody PB, Wong KW, Wang G, Ela W. A review of irregular time series data handling with gated recurrent neural networks. Neurocomputing. 2021;441(7):161–78. doi:10.1016/j.neucom.2021.02.046.
3. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. doi:10.1038/s41597-022-01899-x.
4. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU collaborative research database, a freely available multi-center database for critical care research. Sci Data. 2018;5(1):180178. doi:10.1038/sdata.2018.178.
5. Graves A. Long short-term memory. In: Supervised sequence labelling with recurrent neural networks. Vol. 385. Berlin/Heidelberg, Germany: Springer; 2012. p. 37–45. doi:10.1007/978-3-642-24797-2_4.
6. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8(1):6085. doi:10.1038/s41598-018-24271-9.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000–10.
8. Harerimana G, Kim JW, Jang B. A multi-headed transformer approach for predicting the patient's clinical time-series variables from charted vital signs. IEEE Access. 2022;10(1):105993–6004. doi:10.1109/access.2022.3211334.
9. 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(1):23. doi:10.1186/s12874-023-02138-6.
10. Jung J, Kang S, Choi J, El-Kareh R, Lee H, Kim H. Evaluating the impact of explainable AI on clinicians' decision-making: a study on ICU length of stay prediction. Int J Med Inform. 2025;201(2):105943. doi:10.1016/j.ijmedinf.2025.105943.
11. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Sci Rep. 2022;12(1):7166. doi:10.1038/s41598-022-11012-2.
12. Gupta M, Gallamoza B, Cutrona N, Dhakal P, Poulain R, Beheshti R. An extensive data processing pipeline for MIMIC-IV. Proc Mach Learn Res. 2022;193:311–25.
13. Xu J, Minakova N, Sanchez PO, Riezler S. Early prediction of sepsis using time series forecasting. In: Proceedings of the 2023 IEEE 19th International Conference on e-Science (e-Science); 2023 Oct 9–13; Limassol, Cyprus. p. 1–9. doi:10.1109/e-science58273.2023.10254852.
14. Rosnati M, Fortuin V. MGP-AttTCN: an interpretable machine learning model for the prediction of sepsis. PLoS One. 2021;16(5):e0251248. doi:10.1371/journal.pone.0251248.
15. Li Y, Wang Y. Temporal convolution attention model for sepsis clinical assistant diagnosis prediction. Math Biosci Eng. 2023;20(7):13356–78. doi:10.3934/mbe.2023595.
16. Zhu Y, Cheng Y, Zhang T, Zhang L, Hong X, Wang D, et al. Application of the KA-Transformer model to early sepsis prediction: a hybrid network analysis based on time series data. Discover Appl Sci. 2025;7(3):218. doi:10.1007/s42452-025-06628-8.
17. Wang W, Zuo E, Chen C, Chen C, Zhong J, Yan Z, et al. Efficient time series adaptive representation learning via dynamic routing sparse attention. Pattern Recognit. 2025;158(7):111058. doi:10.1016/j.patcog.2024.111058.

18. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. Sci Rep. 2019;9(1):1879. doi:10.1038/s41598-019-38491-0.
19. Ma L, Gao J, Wang Y, Zhang C, Wang J, Ruan W, et al. AdaCare: explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. Proc AAAI Conf Artif Intell. 2020;34(1):825–32. doi:10.1609/aaai.v34i01.5427.
20. Li Q, Li D, Nie W, Jiao H, Wu Z, Liu A. Temporal and spatial analysis in early sepsis prediction via causal disentanglements. IEEE Trans Knowl Data Eng. 2025;37(8):4860–72. doi:10.1109/tkde.2025.3569584.
21. Yang H, Li J, Zhang C, Sierra AP, Shen B. Predictive model for daily risk alerts in sepsis patients in the ICU: visualization and clinical analysis of risk indicators. Precis Clin Med. 2025;8(1):pbaf003. doi:10.1093/pcmedi/pbaf003.
22. Hayat N, Geras KJ, Shamout FE. MedFuse: multi-modal fusion with clinical time-series data and chest X-Ray images. Proc Mach Learn Res. 2022;182:479–503.
23. Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, et al. Developing a new definition and assessing new clinical criteria for septic shock: for the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016;315(8):775. doi:10.1001/jama.2016.0289.
24. Kamran F, Tjandra D, Heiler A, Virzi J, Singh K, King JE, et al. Evaluation of sepsis prediction models before onset of treatment. NEJM AI. 2024;1(3):AIoa2300032. doi:10.1056/aioa2300032.
25. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271. 2018.
26. Gao J, Xiao C, Wang Y, Tang W, Glass LM, Sun J. StageNet: stage-aware neural networks for health risk prediction. In: Proceedings of The Web Conference 2020; 2020 Apr 20–24; Taipei, Taiwan. p. 530–40. doi:10.1145/3366423.3380136.
27. Wang X, Guo Z, Chai Y, Wang Z, Liao H, Wang Z, et al. Application prospect of the SOFA score and related modification research progress in sepsis. J Clin Med. 2023;12(10):3493. doi:10.3390/jcm12103493.
28. Trzeciak S, Dellinger RP, Chansky ME, Arnold RC, Schorr C, Milcarek B, et al. Serum lactate as apredictor of mortality in patients with infection. Intensive Care Med. 2007;33(6):970–7. doi:10.1007/s00134-007-0563-9.
29. Wang Z, Wang W, Sun C, Li J, Xie S, Xu J, et al. A methodological systematic review of validation and performance of sepsis real-time prediction models. npj Digit Med. 2025;8(1):190. doi:10.1038/s41746-025-01587-1.