ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.2 / 2026 / pp.76-91
https://doi.org/10.69882/adba.chf.2026072

An Explainable Leakage-Audited Framework for Multi-Horizon Sepsis-3 Onset Forecasting from Complex ICU Dynamics

Mohammed Mansour *,1 **and Turker Berk Donmez** α,2

*Mechatronics Engineering Department, Sakarya University of Applied Sciences, University Road, 54050 Serdivan, Sakarya, Türkiye, αBiomedical Engineering Department, Sakarya University of Applied Sciences, University Road, 54050 Serdivan, Sakarya, Türkiye.

ABSTRACT Sepsis-3 onset forecasting remains challenging because event prevalence is low at each patient-time landmark and retrospective electronic health record (EHR) models may inadvertently exploit workflow signals that reflect clinician suspicion rather than underlying patient physiology. We present an explainable, leakage-audited, multi-horizon LightGBM framework for physiology-based Sepsis-3 onset forecasting using MIMIC-IV v3.1. Adult first ICU stays with at least 8 hours of observation were transformed into hourly physiologic trajectories, Sepsis-3 onset labels, and repeated landmark prediction examples. Separate models estimated the probability of Sepsis-3 onset within the next 4, 6, 12, and 24 hours. The final cohort comprised 64,363 ICU stays, including 42,291 forecastable stays and 6,431 forecastable Sepsis-3 cases. The deployment-oriented no-admission-year CLEAN model achieved held-out AUROCs of 0.782, 0.772, 0.750, and 0.720 for the 4-, 6-, 12-, and 24-hour prediction horizons, respectively, with patient-level bootstrap 95% confidence intervals of 0.763–0.801, 0.752–0.792, 0.729–0.769, and 0.699–0.742. A paired leakage audit demonstrated that removing vasopressor exposure, measurement-age channels, and observation-mask channels resulted in AUROC changes of less than 0.01, indicating that predictive performance was primarily driven by physiologic information rather than workflow-related artifacts. Temporal validation on the latest deidentified admission-year group yielded a 6-hour AUROC of 0.762. At the 6-hour horizon, a top-2% no-year alarm policy with a 6-hour cooldown detected 17.8% of sepsis cases within 0–6 hours, 29.3% within 0–24 hours, and 40.3% before onset overall while generating 15.5 alerts per 100 patient-days. SHAP-based explainability analyses consistently identified neurologic status, temperature, urine-output dynamics, renal markers, inflammatory variables, vital signs, and lactate-related dynamics as the dominant predictors across forecasting horizons. These findings provide a robust and explainable physiology-based framework for future studies on Sepsis-3 onset forecasting in complex critical-care systems.

KEYWORDS
Explainable AI
SHAP
Sepsis-3
MIMIC-IV
Dynamic landmark forecasting

INTRODUCTION

Sepsis remains one of the central unsolved problems in acute care because its early clinical manifestations are nonspecific, its trajectory can change over hours, and delayed treatment is associated with preventable organ dysfunction and death. Recent clinical AI studies have moved from retrospective discrimination metrics toward real workflow impact: the COMPOSER deployment study reported that an EHR-integrated deep-learning sepsis predictor was associated with improved sepsis bundle compliance and reduced in-hospital sepsis mortality in two emergency departments (Boussina *et al.* 2024). At the same time, recent systematic reviews show that sepsis prediction models differ widely in target definitions, data windows, model classes, validation strategies, and implementation maturity (Gao *et al.* 2024b; Zubair *et al.* 2025; Zhang *et al.* 2026). These observations motivate a model-development strategy that treats timing, interpretability, and alarm burden as first-class scientific endpoints rather than secondary engineering details.

Large critical-care EHR resources have made reproducible sepsis modeling possible, but they also expose the methodological fragility of retrospective clinical prediction. MIMIC-IV provides deidentified hospital and ICU data from Beth Israel Deaconess Medical Center and is now a standard benchmark for critical-care informatics (Johnson *et al.* 2023). The current MIMIC-IV v3.1 PhysioNet release adds a citable, versioned data object for reproducible research (Johnson *et al.* 2024). Because each version changes the available tables, item definitions, and timestamps that downstream pipelines consume, a sepsis forecasting study must report not only the model family and metrics but also the database version, cohort construction, label windows, and feature provenance.

The recent machine-learning literature supports the potential of EHR-based sepsis prediction, but it also shows that high apparent performance is not sufficient for clinical credibility. A systematic review of machine-learning-based early sepsis prediction using EHRs found that most studies used longitudinal data, but only a

Manuscript received: 4 May 2026,
Revised: 28 June 2026,
Accepted: 10 July 2026.

[1]mohammedmansour@subu.edu.tr (**Corresponding author**)
[2]turkerberkdonmez@yahoo.com

minority reported long prediction horizons or prospective validation, and very few made complete code available (Islam *et al.* 2023). A 2024 network meta-analysis concluded that machine-learning algorithms can be effective for sepsis prediction, while also emphasizing heterogeneity across cohorts and algorithms (Gao *et al.* 2024b). More recent reviews similarly highlight that reported AUROCs can be high while evidence for calibration, external validity, implementation feasibility, and alert management remains incomplete (Zubair *et al.* 2025; Zhang *et al.* 2026).

The clinical endpoint also matters. Many sepsis models predict mortality after sepsis is already recognized, which is useful for prognosis but different from anticipating the onset of Sepsis-3 physiology. Interpretable prognosis studies in MIMIC-IV have shown that tree-based models can identify clinically plausible drivers such as Glasgow Coma Scale, blood urea nitrogen, respiratory rate, urine output, and age (Hu *et al.* 2022). MIMIC-IV mortality prediction studies have also reported strong discrimination with reduced feature sets and ensemble models (Gao *et al.* 2024a). However, mortality prediction among already septic patients does not directly answer the bedside early-warning question: given the current patient state, will Sepsis-3 onset occur in a future time window?

Recent studies have therefore begun to focus on earlier and more operational targets. Liu et al. developed an interpretable MIMIC-IV emergency-triage model for sepsis risk and showed that comprehensive structured triage information outperformed vital signs alone, with Gradient Boosting reaching the highest AUC among the tested models (Liu *et al.* 2025). Duval et al. studied a related anticipatory task, predicting vasopressor initiation in ICU sepsis patients using an interpretable EHR-based design with matched case-control windows, and reported a Random Forest validation AUROC of 0.75 (Duval *et al.* 2025). These studies demonstrate the value of routinely collected physiology, but they also underline the importance of specifying the landmark time, the lookback window, and the event window with enough precision to avoid optimistic retrospective labeling.

Interpretability is not an optional cosmetic layer in this setting. A sepsis early-warning model may influence antibiotics, cultures, fluids, monitoring intensity, and ICU resource allocation; therefore, clinicians need to know whether the prediction is being driven by plausible physiology or by artifacts of documentation and ordering behavior. Recent MIMIC-IV interpretability work has shown that explanation methods can reveal model signals but can also expose demographic reliance and fairness concerns (Meng *et al.* 2022). Disease-specific interpretable studies have used SHAP to present global feature rankings, feature-effect plots, and individual explanations for sepsis prognosis and complications (Hu *et al.* 2022; Lu *et al.* 2022). In a sepsis forecasting study, global SHAP, local SHAP, and feature-dependence plots should therefore be reported alongside discrimination, calibration, and alarm-policy metrics.

A specific risk in sepsis forecasting is clinician-suspicion leakage. Variables such as newly ordered lactate, fresh coagulation tests, cultures, antibiotics, or vasopressor exposure may encode the bedside team's concern before the model observes overt organ dysfunction. If these signals are included without audit, the model may learn to reproduce clinical workflow rather than detect physiology earlier. The problem is particularly relevant for retrospective EHR studies because availability itself is informative: a missing or recently measured laboratory value can function as a proxy for clinical suspicion. Recent operational studies explicitly discuss realistic lookback windows, matching, and alert burden as ways to make retrospective prediction closer to real-time deployment (Duval *et al.* 2025; Boussina *et al.* 2024).

The present study is designed around that gap. We frame the prediction task as dynamic landmark forecasting: at each ICU time τ, the model estimates whether Sepsis-3 onset will occur in the interval $(\tau, \tau + h]$ for horizons $h \in \{4, 6, 12, 24\}$ hours. We use adult first ICU stays from MIMIC-IV v3.1, construct hourly physiologic features and temporal trend features, and train one LightGBM classifier per horizon. We then perform a paired feature-set audit by comparing a FULL model against a CLEAN model that removes the vasopressor flag, measurement-age channels, and observation-mask channels that can proxy clinician suspicion. Because deidentified admission-year group is not a bedside physiologic variable, the deployment-oriented primary analysis uses the no-admission-year CLEAN model, while the year-including CLEAN model is retained as an internal benchmark.

Our contribution is therefore not another single AUROC table, but a reproducible, physiology-focused sepsis forecasting pipeline. In the current MIMIC-IV v3.1 cohort, the no-year CLEAN model reaches held-out test AUROCs of 0.782, 0.772, 0.750, and 0.720 for 4-, 6-, 12-, and 24-hour horizons, respectively. At the 6-hour horizon, the recommended top-2 percent alarm policy with a 6-hour cooldown detects 17.8 percent of forecastable sepsis stays within the horizon-concordant 0–6 hour window, 29.3 percent within 0–24 hours, and 40.3 percent under a broader any-pre-onset definition. SHAP analysis identifies Glasgow Coma Scale total score, body temperature, short-window urine-output variability, and blood urea nitrogen as dominant predictors, aligning the model's behavior with organ dysfunction rather than measurement-process artifacts. This combination of leakage audit, multi-horizon forecasting, alarm-policy evaluation, and patient-level explanation is intended to support transparent analysis of Sepsis-3 early warning.

RELATED WORK

Systematic reviews and reporting gaps

Recent reviews show a rapidly expanding but methodologically heterogeneous sepsis prediction literature. Islam *et al.* (2023) reviewed EHR-based machine-learning and deep-learning studies for early sepsis prediction and found that most papers used longitudinal data, but that prediction windows, feature engineering, missing-data handling, and reporting quality varied substantially. Gao *et al.* (2024b) performed a systematic review and network meta-analysis of machine-learning algorithms in sepsis prediction and reported favorable rankings for deep and boosted approaches, while also emphasizing the need for stronger evaluation and algorithm comparison. Zubair *et al.* (2025) reviewed machine-learning and deep-learning approaches for sepsis diagnosis and highlighted the same translational barriers: EHR heterogeneity, class imbalance, interpretability, and limited real-world validation. Zhang *et al.* (2026) focused specifically on emergency department AI sepsis prediction and reported broad variation in AUROC, model type, feature count, and prediction windows across included studies. Our study follows these reviews by reporting not only AUROC but also AUPRC, Brier score, calibration, alarm burden, and lead time.

MIMIC-IV sepsis models and interpretability

Several closely related MIMIC-IV studies predict prognosis or complications after sepsis has been identified. Hu *et al.* (2022) developed interpretable machine-learning models for early in-hospital mortality prediction in critically ill sepsis patients and found that XGBoost performed best, with SHAP highlighting Glasgow Coma Scale, blood urea nitrogen, respiratory rate, urine output, and

age. Lu *et al.* (2022) developed an interpretable model for sepsis-associated encephalopathy, again using SHAP to connect model outputs to clinically relevant features. Gao *et al.* (2024a) reported a MIMIC-IV sepsis mortality model with a reduced feature set and high AUROC, emphasizing interpretability and feature reduction. Zhang *et al.* (2025) extended this prognosis theme to multiple outcomes, using MIMIC-IV to predict rapid recovery, chronic critical illness, and mortality, with CatBoost as the strongest model and urine output, respiratory rate, and temperature among the most important variables. These studies support the relevance of our top SHAP features, but their targets are prognosis among septic patients rather than pre-onset landmark forecasting.

Early warning, deployment, and positioning of the present study

Early-warning studies are closer to our target because they aim to identify sepsis before or near clinical recognition. Liu *et al.* (2025) used MIMIC-IV emergency triage data and compared vital-sign-only modeling with a broader structured EHR feature set; Gradient Boosting achieved the highest reported AUC and SHAP/LIME were used to explain the model. Boussina *et al.* (2024) evaluated the real-world deployment of COMPOSER and reported that deployment was associated with lower sepsis mortality and improved bundle compliance, making it one of the most important recent implementation studies in the field. Shashikumar *et al.* (2025) later integrated a large language model with COMPOSER to extract context from clinical notes for uncertain cases and reported improvements over the standalone model. Yang *et al.* (2026) proposed a semantic abstraction rule engine to transform structured clinical time series into clinician-readable text and fine-tuned lightweight LLMs for early sepsis prediction. These studies demonstrate growing interest in multimodal and language-model-based prediction, while our work keeps the primary model tabular and explainable to prioritize reproducibility, feature auditability, and fast SHAP computation.

The most direct MIMIC-IV onset comparison is the 2024 IEEE MedAI study by Shan *et al.* (2024), which predicted sepsis onset in ICU patients using multiple machine-learning models and 81 features extracted around ICU admission; XGBoost achieved an AUC of 0.81, and feature selection was used to simplify the model. That work is valuable as an onset-focused benchmark, but it uses an admission-centered feature window rather than repeated landmark forecasts over the ICU stay. Duval *et al.* (2025) addressed a neighboring anticipatory ICU problem by predicting vasopressor initiation several hours before therapy in Sepsis-3 patients, using matched windows and SHAP-based interpretation to reduce temporal and selection bias. Our study differs by directly forecasting Sepsis-3 onset at multiple future horizons, explicitly auditing clinician-suspicion proxy features, and evaluating practical alarm policies with cooldown and lead-time distributions.

Taken together, the recent literature supports four design choices for the present work. First, Sepsis-3 forecasting should be framed as a time-indexed prediction problem rather than a static admission classification task. Second, evaluation should include precision-recall behavior, calibration, lead time, and alert burden because sepsis prevalence at each landmark is low and bedside alerts have operational cost. Third, interpretability should include global and local views, not only a single feature-importance table. Fourth, EHR workflow proxies should be audited because they can inflate retrospective performance. The proposed CLEAN LightGBM pipeline combines these points: it uses MIMIC-IV v3.1, repeated landmarks, multi-horizon labels, a paired leakage-ablation experiment, SHAP explanations, and alarm-policy operating points, positioning the study as a reproducible clinical-ML benchmark rather than a black-box performance comparison.

METHODOLOGY

Study design

This study was designed as a retrospective, observational, dynamic prediction study using deidentified adult intensive care unit (ICU) data from MIMIC-IV v3.1 (Johnson *et al.* 2023, 2024). The objective was not to classify a patient once at admission, but to repeatedly estimate future Sepsis-3 onset risk at discrete ICU landmark times. At every eligible landmark time τ, the model receives only information available at or before τ and outputs the probability that Sepsis-3 onset will occur in a future horizon h. The evaluated horizons were $h \in \{4, 6, 12, 24\}$ hours (Figure 1).

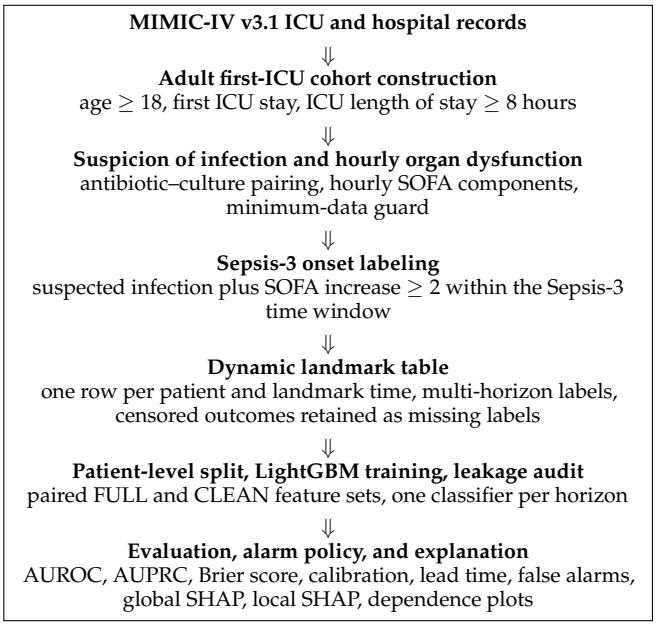

Figure 1 End-to-end methodological workflow. The workflow is intentionally organized so that outcome definitions are created before landmark sampling, patient-level splitting is performed before model fitting, and alarm/explanation analyses are applied only after held-out prediction

Detailed implementation tables, feature definitions, item identifiers, and software versions are provided in the supplementary reproducibility material. The main text retains only the tables needed to interpret cohort construction, model performance, operating points, and explanations.

Data source and cohort construction

The source population was drawn from MIMIC-IV v3.1, a deidentified critical-care database with linked hospital and ICU records (Table 1). The initial cohort was restricted to adult patients aged at least 18 years at ICU admission. Only the first ICU stay per patient was retained to avoid correlation from repeated ICU admissions. ICU stays shorter than 8 hours were excluded because they do not provide enough time for both early physiologic observation and future-event forecasting.

After the base ICU cohort was created, patients whose Sepsis-3 onset occurred within the first 4 hours of ICU admission were not used for forecasting. These patients are clinically important

but do not match the prospective prediction question, because the outcome is already present or nearly present at the first usable landmark. The remaining forecastable cohort contains ICU stays for which a future sepsis, death, or discharge trajectory can be evaluated from repeated landmarks.

Table 1 Cohort construction summary. Counts describe the final analytic pipeline used for the multi-horizon forecasting study

Cohort stage	Count	Interpretation
Adult first ICU stays with ICU length of stay ≥ 8 hours	64,363	Base analytic ICU cohort.
Suspicion-of-infection positive stays	32,521	Antibiotic–culture timing rule satisfied.
Sepsis-3 positive stays in the base cohort	29,004	Suspected infection plus acute organ dysfunction.
Forecastable stays after excluding very-early sepsis	42,291	Dynamic prediction cohort.
Forecastable Sepsis-3 positive stays	6,431	Future-onset sepsis cases available for prediction.
Landmark rows generated for model development	765,510	Repeated patient-time prediction examples.

Suspicion of infection

Suspicion of infection was operationalized as a valid temporal pairing between antimicrobial treatment and microbiologic culture, following the logic used in Sepsis-3 EHR studies (Singer *et al.* 2016; Seymour *et al.* 2016). Culture events were extracted from microbiology records. Antibiotic exposure was extracted from ICU medication administration and hospital pharmacy sources. To reduce under-detection, antimicrobial identification included both structured ICU antibiotic categories and medication-name matching using a broad antibiotic lexicon. This allowed intravenous ICU antibiotics, emergency department medications, oral medications, and other hospital medication records to contribute to the suspected infection signal when they satisfied the required time window.

For a culture-first pattern, the antibiotic had to occur within 24 hours after culture collection. For an antibiotic-first pattern, the culture had to occur within 72 hours after antibiotic initiation. The earliest valid pair was retained as the suspicion-of-infection time. Medication text fields were handled as medication strings rather than numeric-only quantities so that antimicrobial names could be matched reliably even when pharmacy quantity fields had heterogeneous data types.

Hourly SOFA construction and Sepsis-3 endpoint

Hourly Sequential Organ Failure Assessment (SOFA) components were constructed from routinely collected ICU physiology and laboratory data (Table 2). The respiratory component used oxygenation measures and mechanical-ventilation status so that higher respiratory SOFA grades were gated by ventilatory support when appropriate. The cardiovascular component used mean arterial pressure and dose-stratified vasoactive medication exposure. The renal component used serum creatinine and 24-hour urine output. The neurologic, coagulation, and liver components used Glasgow Coma Scale total score, platelet count, and total bilirubin, respectively.

Several harmonization steps were required before SOFA scoring. Fahrenheit temperature charting was converted to Celsius and merged with Celsius temperature charting. Laboratory lookback was allowed to extend to the hospital admission period so that early ICU SOFA did not discard clinically relevant pre-ICU laboratory information. Urine output was aggregated over clinically meaningful rolling windows. Mechanical ventilation and vasoactive medication exposure were mapped from structured ICU events. A minimum-data flag was then applied so that SOFA baselines and peaks were not determined from hours with too few observed organ-system components.

Table 2 Compact Sepsis-3 label algorithm used in the main analysis. The endpoint is an EHR-defined Sepsis-3 onset time, not an independently adjudicated biological state

Label component	Operational definition
Suspicion of infection	Earliest valid antibiotic–culture pair: culture followed by antibiotic within 24 hours, or antibiotic followed by culture within 72 hours. The suspicion time was the earlier timestamp in the valid pair.
SOFA window	Hourly SOFA rows inside $[T_{SI} - 48\text{h}, T_{SI} + 24\text{h}]$. ED and ward laboratory values from hospital admission through ICU time were allowed to support early SOFA reconstruction.
Baseline and acute increase	Baseline SOFA was the minimum hourly SOFA in the Sepsis-3 window; peak SOFA was the maximum. A stay met the organ-dysfunction criterion when peak minus baseline was at least 2 points.
Minimum-data guard	SOFA rows used for baseline/peak selection required at least 3 observed organ-system inputs among MAP, GCS, platelets, bilirubin, creatinine or urine output, and oxygenation.
Onset timestamp	If suspected infection and SOFA increase ≥ 2 were both satisfied, the EHR-defined Sepsis-3 onset time was set to the suspicion-of-infection time.
Predictor exclusion	Antibiotic administrations, cultures, and the suspicion-of-infection timestamp were used only for labels. They were not included as predictors in the landmark models.

Landmark prediction setup and target definition

The prediction problem was formulated using landmarking. For each forecastable ICU stay, prediction rows were emitted at repeated landmark times τ, beginning at ICU hour 4 and continuing every 2 hours until the earlier of ICU hour 72 or the last usable hour before ICU discharge or death. Each row represents the patient's state at a specific time and contains only predictors known at or before that time.

For each horizon $h \in \{4, 6, 12, 24\}$, the binary target was defined as:

$$y_h(\tau) = \begin{cases} 1, & \text{if } \tau < T_{\text{sepsis}} \leq \tau + h, \\ 0, & \text{if the patient is observed through } \tau + h \\ & \text{without Sepsis-3 onset,} \\ \text{missing}, & \text{if discharge or death occurs before } \tau + h \\ & \text{without observed Sepsis-3 onset.} \end{cases}$$

Rows with missing labels for a given horizon were retained in the landmark table but excluded from that horizon's supervised training and metric calculation. This prevents early discharge or death before the prediction window from being incorrectly counted as evidence that sepsis would not have occurred.

Feature engineering

Feature engineering was performed in chronological order to preserve temporal validity (Figure 2). First, current values were created using the most recent observed value available at or before each ICU hour. Second, measurement-age and observation-mask channels were created for the FULL feature set to record how recently each physiologic variable had been observed. Third, short-term trend features were attached at landmark time, including changes over 2, 6, and 12 hours and rolling ranges over 6 and 12 hours. These trend features were computed only from prior values and therefore do not use future information.

The feature vocabulary included demographics, admission metadata, vital signs, neurologic status, routine laboratory tests, renal output, treatment exposure, and temporal dynamics. Sex and hospital admission type were passed to LightGBM as native categorical variables rather than one-hot encoded variables. Deidentified admission-year group was retained only for the year-including internal benchmark; it was removed from the deployment-oriented primary CLEAN model because it may encode secular documentation, coding, ICU practice, case-mix, or COVID-era effects rather than bedside physiology. Missing numeric values were left as missing values in the primary LightGBM model because LightGBM learns default split directions for missingness during tree construction. Because this native handling can still encode subtle measurement-availability information, even after explicit observation masks and measurement-age channels are removed, a median-imputed sensitivity analysis was added in which numeric missing values were replaced using training-set medians and no missingness indicators were supplied.

Patient-level data splitting

The dynamic landmark table was split at the patient level into training, validation, and test partitions using a 70/15/15 ratio. Stratification was based on whether the patient ever developed forecastable Sepsis-3. Because each patient contributes multiple landmark rows, patient-level splitting prevents leakage in which earlier rows from a patient appear in training while later rows from the same patient appear in testing. The training, validation, and test partitions contained 531, 407, 117, 379, and 116, 724 landmark rows, respectively; the full split table is provided in the supplementary material.

LightGBM model development

One LightGBM gradient-boosted decision-tree classifier was trained for each prediction horizon (Ke *et al.* 2017). For a given horizon, only landmarks with observed labels were used for supervised fitting. The same training, validation, and test split assignment was reused across horizons. Hyperparameters were fixed across horizons to make the comparison interpretable: 1000 maximum boosting iterations, learning rate 0.03, 31 leaves, minimum child samples 100, feature fraction 0.8, bagging fraction 0.8, bagging frequency 5, and a fixed random seed. Early stopping used validation AUROC with a patience of 50 boosting rounds.

The 4-, 6-, 12-, and 24-hour models were treated as separate horizon-specific classifiers rather than a single multi-output network. This choice keeps each horizon's class balance, calibration, and alarm operating points explicit. The primary output of each model is a probability $p_h(\tau)$, interpreted as the estimated risk of Sepsis-3 onset within the next h hours conditional on the patient state at τ.

No row over-sampling, negative under-sampling, or focal-loss variant was used for the primary LightGBM models. The models were trained on the observed landmark distribution, and class imbalance was handled at evaluation and deployment stages through AUPRC reporting, validation-selected operating thresholds, and alarm-policy analysis. A regularized logistic-regression baseline used class-balanced weighting as a linear comparator, but this was not part of the primary model specification.

Leakage-ablation audit

Two paired feature sets were evaluated. The FULL feature set included all physiologic values, temporal features, measurement-age channels, observation-mask channels, and the active vasopressor flag. The CLEAN feature set removed 41 clinician-suspicion proxy channels: the vasopressor flag, 20 measurement-age channels, and 20 observation-mask channels. The same model specification was then trained on both feature sets for every horizon.

The leakage audit was defined as the performance difference between FULL and CLEAN models on the held-out test split. If removing workflow-proxy features caused only a small AUROC reduction, the model was interpreted as primarily dependent on physiologic state and trends rather than clinician-suspicion proxies. If the reduction was large, the model would be considered substantially dependent on workflow artifacts. The audit threshold was set before interpretation: a ΔAUROC below 0.02 across horizons would be considered compatible with a conservative CLEAN feature set.

Model evaluation

Discrimination was evaluated using AUROC. Because the positive class is rare at individual landmark times, AUPRC was also reported as a primary practical metric. Calibration and probabilistic accuracy were evaluated using Brier score and decile calibration. All metrics were calculated on the validation and held-out test splits, with the test split used for final reporting.

The full reporting plan includes cohort tables, label observability tables, per-horizon performance tables, calibration plots, receiver-operating characteristic curves, precision-recall curves, leakage-ablation tables, alarm operating-point tables, lead-time summaries, and SHAP explanation figures. This structure is intended to make the study auditable from raw cohort construction through patient-level alarm behavior.

Statistical uncertainty was quantified using a patient-level clustered bootstrap on the held-out test split. Each bootstrap replicate sampled ICU stays with replacement and included all eligible landmarks from the sampled stays, preserving within-patient row correlation. Percentile 95% confidence intervals were calculated

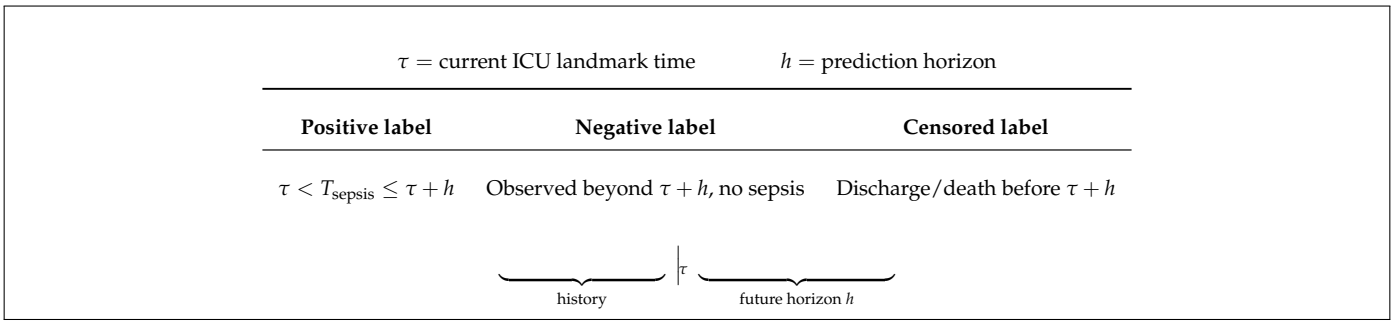

Figure 2 Landmark labeling scheme. The model sees only historical information up to τ, while the target asks whether Sepsis-3 onset occurs in the future interval $(\tau, \tau + h]$.

from 300 bootstrap replicates for AUROC, AUPRC, Brier score, calibration intercept, calibration slope, expected calibration error, and selected alarm-policy metrics.

Additional robustness analyses were defined after the initial internal validation to address concerns about temporal transportability, clinical-score baselines, missingness, lactate-related leakage, SOFA-component overlap, competing events, urine-output outliers, and subgroup performance. Temporal validation trained models on deidentified admission-year groups 2008–2016, selected early stopping on 2017–2019, and tested on 2020–2022 while excluding admission-year group from the feature set. Clinical score baselines included qSOFA, SIRS, current SOFA total, and change in SOFA from the early ICU period. Sensitivity models evaluated median imputation of numeric missing values, removal of lactate-derived features, removal of direct SOFA-component features, a composite sepsis-or-death endpoint, urine-feature winsorization, and subgroup analyses by sex, age group, admission type, deidentified admission-year group, first ICU care unit, and baseline severity.

Alarm-policy analysis

Raw probabilities were converted into alarm policies because a default probability threshold such as 0.5 is unsuitable for a low-prevalence sepsis forecasting problem. Thresholds were selected on the validation split and then applied unchanged to the test split. Candidate thresholds included top-risk row percentiles, validation-set sensitivity targets, and the threshold maximizing validation F1 score.

To mimic practical bedside alert suppression, cooldown windows of 0, 6, and 12 hours were evaluated. When cooldown was active, a patient could not emit another alert until the cooldown interval had elapsed. Alarm performance was summarized at both row level and patient level. Because patient-level first alerts can occur earlier than the nominal model horizon, detection was reported under four nested definitions: horizon-concordant detection, requiring an alert 0–6 hours before onset for the 6-hour model; early-warning detection within 0–24 hours; early-warning detection within 0–48 hours; and any pre-onset alert. Lead time was calculated as onset time minus alert time among detected sepsis stays. Non-sepsis false-alarm rate was defined as the proportion of non-sepsis stays with at least one emitted alarm. Alarm burden was reported as emitted alerts per 100 patient-days.

Explainability analysis

Explainability was performed with Tree SHAP, which computes additive feature attributions for tree-based models (Lundberg *et al.* 2020). Explanations were generated for the no-year CLEAN LightGBM models. For binary LightGBM, SHAP values were interpreted on the model log-odds scale; predicted risks were reported separately after transforming model output to probability. Global SHAP analysis summarized the mean absolute contribution of each feature across a held-out sample of 5,000 labeled landmarks. Because the same patient can contribute multiple landmarks, global SHAP was interpreted as landmark-level model behavior rather than as an independent patient-level causal ranking. Beeswarm plots were used to show attribution magnitude and direction for numeric features. Native categorical variables were interpreted by their SHAP contribution, although their beeswarm colors may appear neutral because categorical levels do not map to a continuous low-to-high color scale; categorical SHAP summaries by level are provided in the supplement.

Local SHAP analysis was performed for a representative held-out sepsis case at a landmark preceding Sepsis-3 onset. This local view decomposes the patient's predicted risk into positive and negative feature contributions at the chosen landmark. A temporal SHAP heatmap was also used for the same patient to show how explanations evolved across earlier landmarks. Finally, dependence plots were produced for the most influential numeric predictors to examine the learned feature-risk relationships.

RESULTS

Cohort yield and event structure

The base analytic cohort contained 64,363 adult first ICU stays with ICU length of stay at least 8 hours. The mean ICU length of stay was 3.57 days and the median ICU length of stay was 1.94 days. The mean age at ICU admission was 64.5 years and the median age was 66 years. Suspicion of infection was identified in 32,521 stays, corresponding to 50.5% of the base cohort. Sepsis-3 criteria were met in 29,004 stays, corresponding to 45.1% of the base cohort and 89.2% of suspicion-of-infection positive stays. After excluding sepsis present at or near ICU admission, 42,291 stays formed the forecastable cohort, of which 6,431 were forecastable Sepsis-3 positive stays. After landmark feasibility filtering, 41,166 stays contributed 765,510 landmark rows to model development. Detailed baseline characteristics by final event type are provided in the supplementary material.

Among landmark-contributing stays, the final event type distribution was 5,374 sepsis stays, 1,645 death-without-forecastable-sepsis stays, and 34,147 discharge-without-forecastable-sepsis stays. Landmark rows were highly imbalanced at every horizon. The observed positive prevalence increased with longer horizon length, from 0.87% at the 4-hour test horizon to 5.87% at the 24-hour test horizon, while censoring also increased with horizon

length because longer windows more often extended beyond discharge or death. The full label-observability table by split and horizon is included in the supplementary material.

Multi-horizon model discrimination and calibration

The deployment-oriented no-year CLEAN LightGBM models showed consistent temporal degradation as the prediction horizon lengthened (Table 3). On the held-out test split, AUROC was 0.782 for 4-hour prediction, 0.772 for 6-hour prediction, 0.750 for 12-hour prediction, and 0.720 for 24-hour prediction. AUPRC increased with horizon length because event prevalence also increased with the longer prediction windows: 0.050 at 4 hours, 0.064 at 6 hours, 0.091 at 12 hours, and 0.145 at 24 hours. Brier scores remained numerically low, reflecting the rare-event probability scale, but increased from 0.0085 at 4 hours to 0.0530 at 24 hours.

Table 3 Deployment-oriented no-year CLEAN LightGBM performance by prediction horizon on the held-out test split. N denotes rows with observed labels for the corresponding horizon

Horizon	Split	N	Positive	Prevalence	AUROC	AUPRC	Brier
4h	Test	110,365	962	0.87%	0.782	0.050	0.0085
6h	Test	106,040	1,450	1.37%	0.772	0.064	0.0131
12h	Test	93,293	2,595	2.78%	0.750	0.091	0.0262
24h	Test	71,818	4,218	5.87%	0.720	0.145	0.0530

Patient-level clustered bootstrap confidence intervals supported the stability of the main discrimination estimates while appropriately accounting for repeated landmarks within patients. At the primary 6-hour horizon, AUROC was 0.772 (95% CI 0.752–0.792), AUPRC was 0.064 (95% CI 0.055–0.077), Brier score was 0.0131 (95% CI 0.0123–0.0142), and calibration slope was 1.10 (95% CI 1.02–1.16). The full bootstrap table is included in the supplementary material (Figure 3).

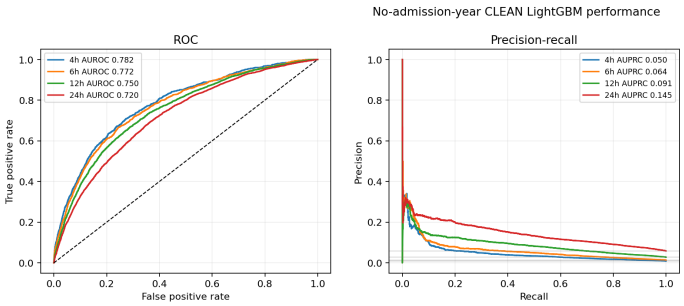

Figure 3 Held-out ROC and precision–recall curves for the no-year CLEAN models across all prediction horizons. Precision–recall panels include prevalence baselines, which are essential for interpreting rare-event prediction. Longer horizons have higher event prevalence but lower AUROC, consistent with the increasing uncertainty of farther-ahead onset prediction

At the 6-hour horizon, decile calibration showed a monotonic increase in observed event frequency as predicted risk increased (Table 4). The lowest decile had mean predicted risk 0.0035 and observed frequency 0.0015, while the highest decile had mean predicted risk 0.0500 and observed frequency 0.0560 (Figure 4). Thus, the model concentrated most sepsis events in the upper risk tail while preserving an interpretable probability scale. The full decile-calibration table is provided in the supplementary material.

Table 4 Held-out test performance with patient-level clustered bootstrap 95% confidence intervals.

Metric	4 h	6 h	12 h	24 h
Prevalence	0.87%	1.37%	2.78%	5.87%
AUROC	0.782 (0.763–0.801)	0.772 (0.752–0.792)	0.750 (0.729–0.769)	0.720 (0.699–0.742)
AUPRC	0.050 (0.042–0.061)	0.064 (0.055–0.077)	0.091 (0.079–0.105)	0.145 (0.128–0.165)
Brier	0.0085 (0.0078–0.0090)	0.0131 (0.0123–0.0142)	0.0262 (0.0241–0.0282)	0.0530 (0.0485–0.0577)
Slope	1.00 (0.94–1.06)	1.10 (1.02–1.16)	1.07 (0.98–1.16)	1.06 (0.95–1.17)
Intercept	−0.01 (−0.19, 0.25)	0.37 (0.06, 0.60)	0.21 (−0.08, 0.50)	0.09 (−0.17, 0.34)
ECE10	0.0007 (0.0006–0.0015)	0.0020 (0.0012–0.0030)	0.0028 (0.0018–0.0048)	0.0037 (0.0029–0.0089)

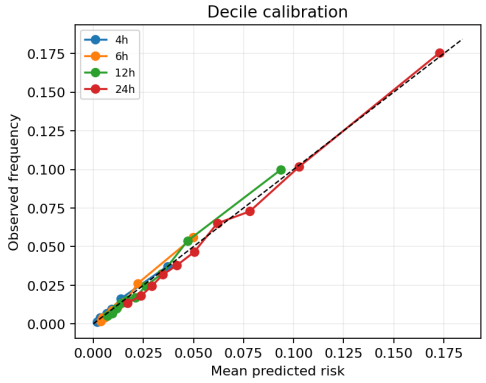

Figure 4 Decile calibration for the no-year CLEAN models across all prediction horizons. Points compare mean predicted risk with observed event frequency in each decile

Baseline, temporal-validation, and sensitivity analyses

The no-year CLEAN LightGBM model outperformed reconstructed clinical scores and a regularized logistic-regression comparator at the primary 6-hour horizon. qSOFA, SIRS, current SOFA total, and early SOFA change had AUROCs between 0.567 and 0.628 (Table 5). A class-weighted regularized logistic model using the CLEAN feature set reached AUROC 0.699 and AUPRC 0.031; because its raw probabilities were poorly calibrated, Platt calibration was fitted on the validation split before reporting Brier score. In contrast, the no-year CLEAN LightGBM model reached AUROC 0.772 and AUPRC 0.064 at the same horizon. This supports the added value of nonlinear tabular modeling and temporal feature interactions beyond simple score thresholds or a linear model.

Table 5 Primary 6-hour model compared with clinical-score and machine-learning baselines on the held-out test split. Brier score is not reported for raw ordinal scores because they are not calibrated probabilities

Comparator	AUROC	AUPRC	Brier	Notes
qSOFA	0.608	0.019	N/A	Ordinal clinical score
SIRS	0.567	0.017	N/A	Ordinal clinical score
Current SOFA total	0.614	0.019	N/A	EHR-reconstructed SOFA at landmark
SOFA change from early ICU period	0.628	0.022	N/A	Early-delta clinical score
Regularized logistic regression	0.699	0.031	0.0134	Platt-calibrated linear CLEAN-feature model
No-year CLEAN LightGBM	0.772	0.064	0.0131	Deployment-oriented nonlinear model

Temporal validation produced slightly lower, but still coherent, discrimination when models were trained on earlier deidentified admission-year groups, validated on 2017–2019, and tested on 2020–2022 after removing admission-year group from the predictors (Table 6). At the 6-hour horizon, temporal-test AUROC

Table 6 Temporal validation using earlier deidentified admission-year groups for training, 2017–2019 for validation, and 2020–2022 for testing. Admission-year group was excluded from temporal-validation predictors. Confidence intervals are patient-level clustered bootstrap intervals.

Metric	4 h	6 h	12 h	24 h
Temporal test *N*	133,859	129,579	116,908	94,774
Positive	1,246	1,897	3,598	6,177
Prevalence	0.93%	1.46%	3.08%	6.52%
AUROC (95% CI)	0.776 (0.759–0.791)	0.762 (0.746–0.777)	0.739 (0.722–0.756)	0.697 (0.679–0.715)
AUPRC (95% CI)	0.034 (0.030–0.040)	0.047 (0.042–0.052)	0.084 (0.076–0.094)	0.139 (0.126–0.152)
Brier (95% CI)	0.0093 (0.0087–0.0099)	0.0144 (0.0135–0.0153)	0.0292 (0.0272–0.0314)	0.0591 (0.0549–0.0628)

was 0.762 (95% CI 0.746–0.777), AUPRC was 0.047 (95% CI 0.042–0.052), and Brier score was 0.0144 (95% CI 0.0135–0.0153). The 24-hour temporal AUROC was 0.697 (95% CI 0.679–0.715), indicating greater degradation for farther-ahead prediction under temporal shift. These results are weaker than random patient-level testing, especially for AUPRC, but support the manuscript's positioning as an internally and temporally validated single-database benchmark rather than an externally validated deployment model.

Temporal alarm-policy behavior was also evaluated for the 6-hour no-year model. With a top-2% validation threshold and 6-hour cooldown, temporal-test horizon-concordant detection was 15.1%, 0–24 hour detection was 28.9%, and any-pre-onset detection was 46.3%. The temporal false-alarm rate was higher than in the random test split at 21.4%, with 17.6 alerts per 100 patient-days. Patient PPV was 9.2% for horizon-concordant detection and 17.7% for 0–24 hour detection, indicating that operating thresholds would need site- and period-specific monitoring before deployment. The full temporal alarm table is provided in the supplementary material.

Sensitivity analyses addressed residual reviewer concerns about missingness, lactate, SOFA-component overlap, deidentified admission-year metadata, competing events, and urine-output outliers (Table 7). At the 6-hour horizon, the year-including internal benchmark reached AUROC 0.779, while the no-year primary model reached AUROC 0.772. Median-imputing all numeric missing values in the CLEAN model reduced AUROC to 0.766, suggesting that native LightGBM missingness handling contributes modestly but does not dominate performance. Removing lactate-derived features produced AUROC 0.776, indicating that the model does not rely strongly on lactate availability or lactate values. Removing direct SOFA-component features produced a larger but still moderate reduction to AUROC 0.758, consistent with the model using early organ-dysfunction patterns that precede formal Sepsis-3 threshold crossing. A no-year urine-winsorized model, which clipped urine-related features at the training-set 1st–99th percentiles, reached AUROC 0.774 (Table 8). A no-year composite endpoint model for Sepsis-3 onset or death within 6 hours reached AUROC 0.808 and AUPRC 0.140, showing that the framework also captures broader deterioration risk but that this is a different clinical endpoint.

Subgroup analysis at the 6-hour horizon showed broadly similar discrimination across sex, age, and admission-year groups, but it also identified heterogeneity that would require prospective monitoring before deployment. AUROC was 0.770 in female patients and 0.786 in male patients. Age-stratified AUROC declined from 0.824 in patients aged 18–49 years to 0.751 in patients aged 75 years or older. Deidentified admission-year group performance varied from 0.725 in 2014–2016 to 0.821 in 2020–2022. The full subgroup table, including admission type, ICU type, baseline-severity group, calibration, and AUPRC, is provided in the supplementary material.

Table 7 Reviewer-directed 6-hour sensitivity analyses on the held-out test split. The sepsis-or-death row uses a retrained composite endpoint model and should not be compared as the same target. Full per-horizon sensitivity results are included in the supplementary material

6-hour feature sensitivity	Features	AUROC	AUPRC	Brier
No-year CLEAN LightGBM	129	0.772	0.064	0.0131
Year-including internal benchmark	130	0.779	0.073	0.0131
Median-imputed numeric missing values	130	0.766	0.066	0.0131
No lactate-derived features	124	0.776	0.074	0.0131
No direct SOFA-component features	88	0.758	0.069	0.0131
No-year urine winsorization	129	0.774	0.064	0.0132
No-year sepsis-or-death endpoint	129	0.808	0.140	0.0161

Table 8 Leakage-ablation discrimination results on the held-out test split. Negative deltas indicate lower CLEAN performance relative to FULL. Confidence intervals for ΔAUROC are patient-level paired bootstrap intervals. Calibration metrics are reported separately in Table 9.

Metric	4 h	6 h	12 h	24 h
AUROC (FULL)	0.793	0.784	0.757	0.731
AUROC (CLEAN)	0.785	0.779	0.755	0.728
ΔAUROC (95% CI)	−0.008 (−0.014, −0.002)	−0.005 (−0.010, +0.001)	−0.003 (−0.007, +0.002)	−0.002 (−0.007, +0.002)
AUPRC (FULL)	0.066	0.077	0.104	0.163
AUPRC (CLEAN)	0.057	0.073	0.099	0.154
ΔAUPRC	−0.009	−0.004	−0.005	−0.009

Subgroup alarm summaries showed similar heterogeneity in operating behavior. For the top-2% 6-hour policy, 0–24 hour detection was 22.7% in female patients and 30.2% in male patients, while non-sepsis false-alarm rates were 13.6% and 15.5%, respectively. Across age groups, 0–24 hour detection was highest in patients aged 18–49 years (30.7%) and lowest in patients aged 75 years or older (23.8%); the oldest group also had the highest non-sepsis false-alarm rate (17.7%). Across deidentified admission-year groups, 0–24 hour detection ranged from 22.3% in 2014–2016 to 29.8% in 2011–2013, and alerts per 100 patient-days ranged from 10.6 to 18.9. These differences do not invalidate the aggregate model, but they indicate that subgroup calibration and alert burden would need active monitoring before prospective use.

Leakage-ablation results

The leakage-ablation analysis compared the FULL feature set against the CLEAN feature set after removing clinician-suspicion and treatment-proxy channels (Table 9). Across all horizons, the AUROC reduction after removing these channels was small: -0.008 at 4 hours, -0.005 at 6 hours, -0.003 at 12 hours, and -0.002 at 24 hours. Patient-level paired bootstrap intervals for the AUROC deltas were narrow and all centered near zero. The largest AUPRC reduction was -0.009. These changes are below the 0.02 AUROC concern threshold, supporting the interpretation that the model's performance is largely driven by physiology and prior physiologic trends rather than by measurement-process proxies. This leakage audit isolates workflow-proxy removal; the no-year primary model then additionally removes deidentified admission-year metadata for deployment-oriented reporting.

Table 9 Leakage-ablation calibration results on the held-out test split. All Brier-score increases after CLEAN feature filtering are below 3×10^{-4}, so probability calibration is essentially preserved

Horizon	Brier FULL	Brier CLEAN	ΔBrier
4h	0.00839	0.00844	+0.00004
6h	0.01306	0.01309	+0.00004
12h	0.02603	0.02610	+0.00007
24h	0.05250	0.05274	+0.00024

Alarm-policy results

The 6-hour model was selected as the primary operating horizon because it provides a clinically interpretable risk-enrichment window while preserving adequate positive-event density. At the 6-hour horizon, the no-year top-2% alarm policy with 6-hour cooldown detected 17.8% of sepsis cases within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under a broader any-pre-onset definition. The same policy produced a 16.5% non-sepsis false-alarm rate and 15.50 alerts per 100 patient-days (Figure 5). Because any-pre-onset detection can include alerts earlier than the nominal 6-hour horizon, the horizon-concordant and early-warning definitions are the primary clinical interpretation.

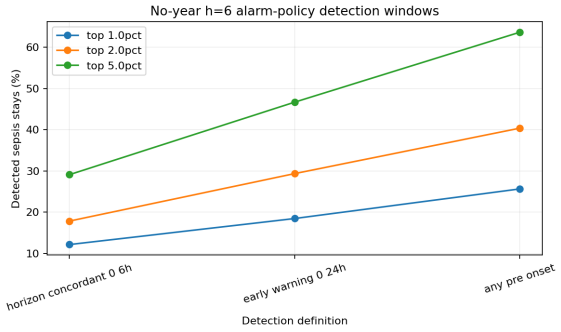

Figure 5 Alarm-policy detection for the 6-hour no-year CLEAN model under alternative lead-time definitions. The same validation-derived threshold and 6-hour cooldown can be interpreted as horizon-concordant detection (0–6h before onset), broader early-warning detection (0–24h or 0–48h), or any pre-onset enrichment

The revised detection-window analysis clarifies the relationship between a 6-hour prediction model and patient-level alert timing (Table 10). For the top-2% no-year policy, horizon-concordant detection within 0–6 hours before onset was 17.8%, with a median lead time of 2.34 hours. The same policy detected 29.3% of sepsis stays within a broader 0–24 hour early-warning window and 32.8% within 0–48 hours. The 40.3% detection corresponds to the broad any-pre-onset definition, which is useful for risk enrichment but should not be interpreted as strictly horizon-concordant 6-hour detection. The denominator distinction is important: the h=6 test split contained 106,040 observed landmark rows and 1,450 positive landmark rows, but patient-level alarm detection used 808 sepsis-positive ICU stays as the denominator. The top-2% policy alerted 1,221 ICU stays and emitted 1,508 cooldown-filtered alerts; this corresponds to 10.5 alerts per horizon-concordant detected case and Patient PPV of 11.8% under the strict 0–6 hour definition.

Table 10 No-year h=6 alarm-policy analysis under horizon-concordant, early-warning, and broad any-pre-onset definitions. Patient PPV is calculated as detected sepsis stays divided by all alerted stays for that definition

Policy	Detection definition	Caught / 808	Detection	Median lead	Patient PPV	False alarm	Alerts / 100 pt-d
Top 1%	Horizon-concordant 0–6h	98	12.1%	2.35 h	13.8%	9.3%	8.21
Top 1%	Early-warning 0–24h	149	18.4%	4.00 h	20.9%	9.3%	8.21
Top 1%	Any pre-onset	207	25.6%	7.63 h	29.1%	9.3%	8.21
Top 2%	Horizon-concordant 0–6h	144	17.8%	2.34 h	11.8%	16.5%	15.50
Top 2%	Early-warning 0–24h	237	29.3%	4.87 h	19.4%	16.5%	15.50
Top 2%	Any pre-onset	326	40.3%	9.44 h	26.7%	16.5%	15.50
Top 5%	Horizon-concordant 0–6h	235	29.1%	2.58 h	10.7%	31.1%	34.16
Top 5%	Early-warning 0–24h	377	46.7%	5.83 h	17.1%	31.1%	34.16
Top 5%	Any pre-onset	514	63.6%	11.41 h	23.4%	31.1%	34.16

Global SHAP analysis

Global SHAP analysis of the 6-hour no-year CLEAN model identified neurologic status, temperature, admission route, renal-output dynamics, renal laboratory values, inflammatory markers, acid-base dynamics, and vital signs as the strongest contributors (Table 11). Glasgow Coma Scale total score was the dominant predictor by mean absolute SHAP value. Body temperature, hospital admission type, range of 6-hour urine output over the previous 12 hours, blood urea nitrogen, white blood cell count, bicarbonate change, respiratory rate, and heart rate followed. Categorical variables such as hospital admission type are shown with neutral coloring in beeswarm visualizations because they are native categorical LightGBM inputs rather than ordered continuous variables; a separate categorical SHAP summary by level is included in the supplementary material.

SHAP rankings were interpreted as model-attribution summaries rather than causal rankings (Figure 7). Several predictors are physiologically correlated, including blood urea nitrogen, creatinine, urine output, and SOFA-related renal dysfunction; similarly, heart rate, respiratory rate, temperature, and white blood cell count can co-vary during systemic inflammation. Correlation can distribute or shift SHAP credit among related variables, so global rankings were interpreted together with dependence plots, local explanations, and clinical domain knowledge.

Local SHAP example

A representative held-out sepsis case was selected for local explanation (Table 12). At ICU hour $\tau = 8$, the 6-hour no-year model assigned a predicted Sepsis-3 onset risk of 0.452 (Figure 7). The patient's Sepsis-3 onset occurred 4.35 hours after the landmark, placing the event inside the 6-hour prediction window. The largest positive contribution was Glasgow Coma Scale total score of 3, followed by hemoglobin 14.7 g/dL, body temperature 37.56 °C, lactate 3.5 mmol/L, 6-hour lactate range, emergency admission type, blood glucose 174 mg/dL, total bilirubin 0.6 mg/dL, white blood cell count 20.2 K/uL, heart-rate range, urine-output variability, respiratory rate, and heart rate. The explanation is compatible with the patient's bedside pattern: depressed neurologic status, elevated lactate, leukocytosis, borderline hypotension, tachycardia, and stress hyperglycemia all push the prediction toward imminent Sepsis-3 onset. The locally large hemoglobin contribution should be interpreted as context-dependent model attribution, not as an isolated causal claim; correlated severity, fluid status, sampling context, and interactions with other features can shift SHAP credit for a single patient (Figure 8).

Table 11 Top 20 global SHAP features for the 6-hour no-year CLEAN model. Values are mean absolute SHAP contributions on the LightGBM log-odds scale across a held-out landmark sample

Rank	Feature	Mean |SHAP|
1	Glasgow Coma Scale total score	0.309
2	Body temperature	0.123
3	Hospital admission type	0.089
4	Range of 6-hour urine output over previous 12 hours	0.089
5	Blood urea nitrogen	0.076
6	White blood cell count	0.063
7	Change in serum bicarbonate over previous 12 hours	0.057
8	Respiratory rate	0.049
9	Heart rate	0.048
10	Change in hemoglobin over previous 12 hours	0.046
11	Change in white blood cell count over previous 6 hours	0.038
12	Change in serum creatinine over previous 12 hours	0.033
13	Hemoglobin	0.033
14	Total bilirubin	0.032
15	Change in white blood cell count over previous 12 hours	0.030
16	Change in peripheral oxygen saturation over previous 6 hours	0.030
17	Range of lactate over previous 6 hours	0.029
18	Age at ICU admission	0.029
19	Range of lactate over previous 12 hours	0.027
20	Serum sodium	0.025

Table 12 Top local SHAP contributors for a held-out sepsis patient at ICU hour 8. The no-year model predicted 45.2% risk of Sepsis-3 onset within 6 hours; onset occurred 4.35 hours later. SHAP values are log-odds contributions

Rank	Feature	Value at $\tau = 8$h	SHAP	|SHAP|
1	Glasgow Coma Scale total score	3.0	1.016	1.016
2	Hemoglobin	14.7 g/dL	0.336	0.336
3	Body temperature	37.56 °C	0.278	0.278
4	Lactate	3.5 mmol/L	0.276	0.276
5	Range of lactate over previous 6 hours	3.5 mmol/L	0.223	0.223
6	Hospital admission type	Emergency ward admission	0.152	0.152
7	Blood glucose	174 mg/dL	0.149	0.149
8	Total bilirubin	0.6 mg/dL	0.145	0.145
9	White blood cell count	20.2 K/uL	0.133	0.133
10	Range of heart rate over previous 12 hours	3 beats/min	0.127	0.127
11	Range of 6-hour urine output over previous 12 hours	150 mL	0.117	0.117
12	Respiratory rate	22 breaths/min	0.116	0.116
13	Heart rate	105 beats/min	0.099	0.099
14	Range of respiratory rate over previous 12 hours	1 breath/min	0.096	0.096
15	Urine output in previous 6 hours	116 mL	0.092	0.092

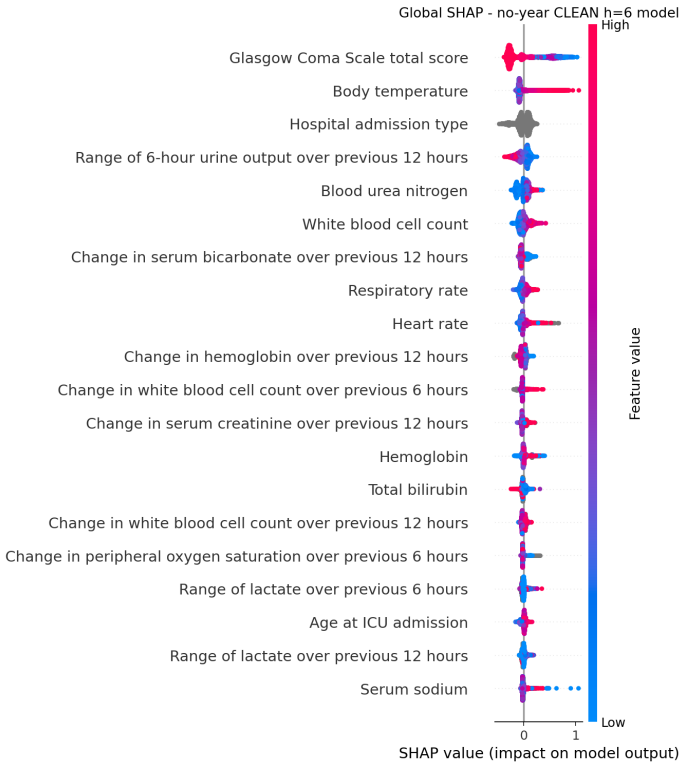

Figure 6 Global SHAP beeswarm plot for the primary 6-hour no-year CLEAN model. Other horizon-specific beeswarm plots and categorical SHAP summaries are provided in the supplementary figure set. SHAP values are on the model log-odds scale

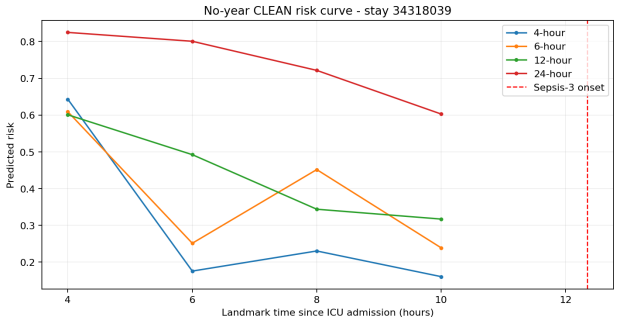

Figure 7 Patient-level 6-hour risk curve for the representative held-out sepsis case. The selected landmark is ICU hour 8, and Sepsis-3 onset occurred 4.35 hours later.

Dependence analysis of the four most influential numeric variables

Dependence plots were generated for the four most influential numeric predictors in the 6-hour no-year CLEAN model: Glasgow Coma Scale total score, body temperature, range of 6-hour urine output over the previous 12 hours, and blood urea nitrogen (Table 13). Lower Glasgow Coma Scale scores strongly increased risk, while a normal score of 15 was protective. Higher temperature values, especially above approximately 37.4 °C in the held-out distribution, increased risk. Lower 12-hour urine-output variability increased risk, consistent with persistently low or nonrecovering urine output, whereas very high variability was associated with lower SHAP values. Blood urea nitrogen showed a broadly

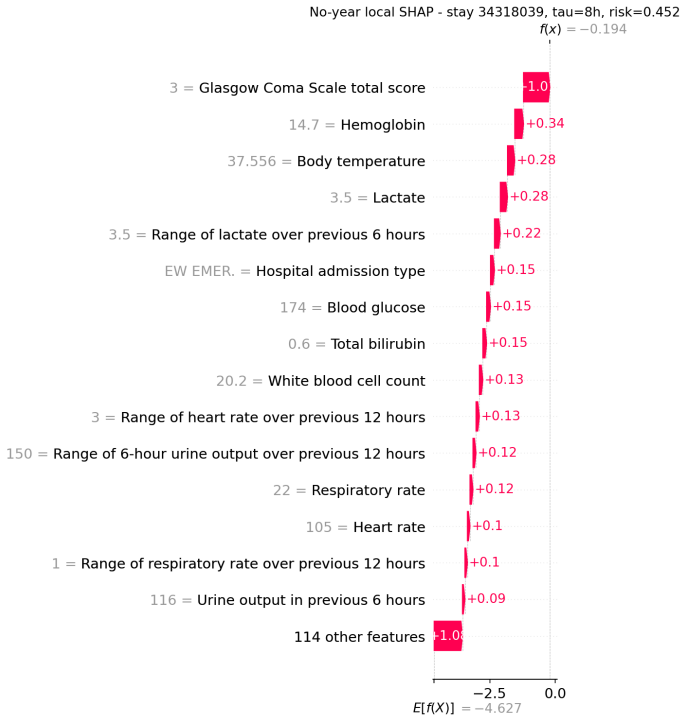

Figure 8 Local SHAP waterfall for the representative held-out sepsis case at ICU hour 8. The waterfall decomposes the selected landmark prediction on the LightGBM log-odds scale; predicted probability is reported separately in the text and table. The temporal SHAP heatmap is provided in the supplementary figure set

monotonic risk increase, with high values carrying positive SHAP contributions.

Table 13 Dependence-plot summary for the four most influential numeric SHAP features in the 6-hour CLEAN model. Ranges correspond to empirical bins from held-out landmarks. Per-feature clinical interpretations are given in the surrounding paragraph; the full quantitative summary with all 10 deciles per feature is provided in the supplementary material

Feature	Low / reference range	Mean SHAP	High-risk range	Mean SHAP
Glasgow Coma Scale total score	3–7 points	+0.649	15 points	−0.262
Body temperature	36.5–36.6 °C	−0.090	37.44–39.89 °C	+0.512
6 h urine output range over prior 12 h	0–99 mL	+0.112	1,560–4,258 mL	−0.255
Blood urea nitrogen	1–7 mg/dL	−0.098	48–181 mg/dL	+0.149

Summary of principal findings

The results support four main conclusions (Figure 9). First, Sepsis-3 onset forecasting is feasible from routinely collected ICU physiology, but the task remains difficult because landmark-level prevalence is low and farther horizons are less discriminable. Second, the CLEAN model retains almost all FULL-model discrimination after removal of measurement-process and treatment-proxy channels, reducing concern that performance is driven by clinician-suspicion leakage. Third, the 6-hour no-year model provides a risk-enrichment operating horizon only when detection windows

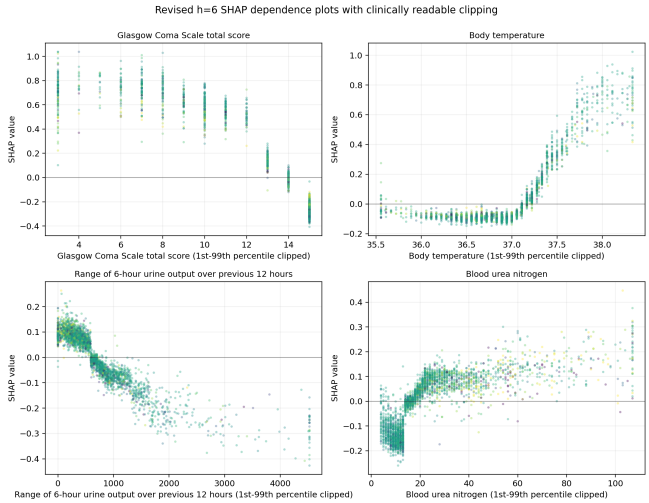

Figure 9 Revised SHAP dependence plots for the four most influential numeric predictors at the 6-hour horizon. Axes include clinical units and are clipped to the 1st–99th percentile for readability, reducing distortion from extreme urine-output values while preserving the learned nonlinear relationship between feature value and log-odds contribution to predicted Sepsis-3 risk

are stated explicitly: the top-2% alarm policy detects 17.8% of test sepsis stays within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under a broader any-pre-onset definition. Fourth, SHAP explanations are interpretable at both global and local levels, with dominant contributions from neurologic status, temperature, renal-output dynamics, blood urea nitrogen, inflammatory markers, vital signs, and lactate-related dynamics.

DISCUSSION

Principal interpretation

This study shows that Sepsis-3 onset can be forecast from routinely collected ICU physiology using a leakage-audited, multi-horizon tabular model, but it also shows why this task should not be interpreted like a static diagnosis or mortality classification problem. The deployment-oriented no-year CLEAN LightGBM model achieved held-out AUROCs of 0.782, 0.772, 0.750, and 0.720 for 4-, 6-, 12-, and 24-hour horizons, respectively. This pattern is clinically and statistically expected: as the horizon lengthens, the model is asked to predict a more remote and more uncertain physiologic transition. The fact that AUROC decreases with longer horizons while AUPRC increases reflects two simultaneous forces: farther-ahead prediction is harder, but longer windows contain more positive events and therefore have higher observed prevalence.

The main methodological finding is that removing clinician-suspicion proxy features produced only minimal performance loss. The largest AUROC difference between FULL and CLEAN models was -0.008 at the 4-hour horizon, and the 6-hour model lost only -0.005 AUROC. This is important because retrospective EHR sepsis prediction is vulnerable to inflated performance when models exploit signals such as recently ordered lactate, recent coagulation tests, culture ordering, or vasopressor exposure. In this pipeline, measurement-age channels, observation-mask channels, and the active vasopressor flag were removed from the primary model, yet performance remained stable. The resulting interpretation is that

the model is learning from physiologic state and prior physiologic dynamics rather than simply reproducing bedside workflow.

Comparison with model-performance literature

The performance level of the present model is best understood in relation to the exact prediction target. Studies that predict mortality after sepsis has already been identified often report higher discrimination than our onset-forecasting model. For example, Hu *et al.* (2022) reported an XGBoost AUC of 0.884 for early in-hospital mortality prediction among ICU patients with sepsis, and Gao *et al.* (2024a) reported a very high AUROC for sepsis mortality prediction using a reduced MIMIC-IV feature set. These tasks are clinically valuable, but they are not equivalent to predicting future Sepsis-3 onset in a mixed ICU population. Mortality models can use information from an already septic cohort, often after substantial organ dysfunction has appeared. Our target is earlier and stricter: at each landmark, the patient may not yet satisfy Sepsis-3 criteria, and the model must estimate whether onset will occur in a future window.

Compared with studies that are closer to onset prediction, our results are in a plausible range. Liu *et al.* (2025) developed an interpretable MIMIC-IV emergency-triage sepsis model and reported that Gradient Boosting achieved an AUC of 0.83 when comprehensive triage information was used. Shan *et al.* (2024) predicted sepsis onset in ICU patients using MIMIC-IV and reported an XGBoost AUC of 0.81 using admission-centered features. Our 4- and 6-hour AUROCs are slightly lower than these values, but our design differs in ways that make the task more conservative: repeated landmark prediction across the ICU stay, censoring-aware horizon labels, patient-level splitting, removal of workflow-proxy channels, and exclusion of deidentified admission-year group from the primary model. Duval *et al.* (2025) predicted vasopressor initiation in Sepsis-3 patients several hours before therapy and reported a Random Forest validation AUROC of 0.75; our 6-hour AUROC of 0.772 is similar in magnitude but targets Sepsis-3 onset rather than hemodynamic intervention.

The supplementary material includes a structured comparison table summarizing these related studies and the limits of direct metric comparison.

These comparisons reinforce a key point emphasized by recent systematic reviews: sepsis prediction studies are difficult to compare unless the target, prediction window, data source, censoring rule, and implementation setting are aligned (Islam *et al.* 2023; Gao *et al.* 2024b; Zubair *et al.* 2025; Zhang *et al.* 2026). A model with a higher AUROC may be solving a less prospective task, using a richer or less controlled feature set, or predicting a more downstream endpoint. For that reason, the most meaningful contribution of the present work is not that it outperforms every previous model numerically; rather, it offers a transparent, leakage-audited benchmark for dynamic Sepsis-3 onset forecasting.

The added baseline analyses further clarify the model's contribution. At the 6-hour horizon, qSOFA, SIRS, current SOFA, and early SOFA change achieved AUROCs between 0.567 and 0.628, while regularized logistic regression using the same CLEAN feature set reached AUROC 0.699. The no-year CLEAN LightGBM model reached AUROC 0.772 and AUPRC 0.064. This does not establish algorithmic novelty, but it does show that nonlinear tree-based modeling and engineered temporal features add measurable value over simpler clinical scores and a linear comparator under the same landmark target.

Interpretation of the horizon-specific results

The 4-hour model produced the highest AUROC, which is expected because near-term onset is closer to the current physiologic state. However, the 4-hour label was extremely rare at the landmark level, with only 0.87% observed test prevalence. This scarcity limits AUPRC and makes fixed high-probability thresholds unrealistic. The 24-hour model had lower AUROC but higher AUPRC because positive labels were more frequent across the longer window. The 6-hour horizon provided the most balanced operating point: it retained near-term discrimination while giving enough time for clinical reassessment, cultures, antibiotics, fluids, closer monitoring, or escalation.

The practical alarm results support this choice, but only when the detection window is stated explicitly. At the h=6 top-2% operating point with a 6-hour cooldown, the no-year model detected 17.8% of sepsis stays within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under the broad any-pre-onset definition, with 15.5 alerts per 100 patient-days. This is better interpreted as patient risk enrichment than as high-sensitivity bedside detection. The top-5% policy detected more patients but nearly doubled the false-alarm rate and alert burden. The top-1% policy had lower burden but missed more cases. This is consistent with the implementation literature: early-warning models can only help if alert rates are acceptable and the receiving clinical team trusts the signal (Boussina *et al.* 2024).

The revised alarm-window analysis also changes the interpretation of lead time. The broad any-pre-onset definition is useful for assessing patient-level enrichment, but it can include alerts much earlier than the nominal 6-hour target. Under the stricter horizon-concordant 0–6 hour definition, the top-2% policy detected 17.8% of sepsis stays with a median lead time of 2.34 hours. Under a broader but clinically plausible 0–24 hour early-warning definition, detection was 29.3% with a median lead time of 4.87 hours. Thus, the alarm analysis should be read as a family of operating definitions rather than a single detection number. The any-pre-onset result is secondary and should not be described as pure 6-hour horizon performance.

Clinical meaning of the leakage-ablation result

The CLEAN versus FULL comparison is one of the central findings of the study. In many EHR models, missingness and measurement recency are not merely technical nuisances; they can encode clinical suspicion. A newly obtained lactate, coagulation test, culture, or arterial blood gas often means that a clinician is already worried. If such process information is used without qualification, the model may appear predictive while functioning as a detector of clinician concern. The present CLEAN model intentionally removes measurement-age and observation-mask features from the primary model. The minimal performance loss suggests that the remaining physiologic signals carry genuine predictive information.

This finding has two implications. First, it strengthens the scientific validity of the reported SHAP explanations, because the top variables are actual physiology and temporal trends rather than test-ordering artifacts. Second, it makes the feature-set rationale more transparent. The study can report that a more permissive feature set was tested, that suspected leakage channels were removed, and that the core conclusions did not depend on those channels. This directly addresses concerns raised in systematic reviews about heterogeneity, incomplete reporting, and limited reproducibility in sepsis prediction studies (Islam *et al.* 2023; Zubair *et al.* 2025).

Residual missingness leakage cannot be fully excluded, because native LightGBM split directions can still learn whether a nu-

meric value is missing even when explicit observation masks and measurement-age variables are removed. The median-imputation sensitivity analysis addresses this concern directly: forcing numeric missing values to training-set medians reduced 6-hour AUROC to 0.766. This modest drop relative to the no-year primary model (0.772) suggests that implicit missingness contributes some signal, but the model's discrimination is not predominantly a missingness artifact. Similarly, removing lactate-derived features left 6-hour AUROC at 0.776, reducing concern that performance is driven by lactate ordering or lactate-specific suspicion pathways.

Discussion of global SHAP findings

The global SHAP profile is consistent with prior interpretable sepsis studies. Glasgow Coma Scale total score was the strongest global feature in the 6-hour model. This aligns with Hu *et al.* (2022), who also identified GCS as the most important feature in an interpretable MIMIC-IV sepsis mortality model. Although our endpoint is future Sepsis-3 onset rather than mortality, severe neurologic depression is a marker of systemic illness, hypoperfusion, encephalopathy, sedation burden, or respiratory failure. A low GCS may therefore signal either organ dysfunction already emerging or vulnerability to rapid deterioration.

Blood urea nitrogen was another major contributor, and its dependence plot showed increasing SHAP contribution at higher values. This agrees with Hu *et al.* (2022), who identified blood urea nitrogen among the top predictors of sepsis prognosis. BUN is not a pure sepsis biomarker; it reflects renal perfusion, catabolic stress, dehydration, renal dysfunction, and critical illness severity. In our model, BUN likely contributes because renal dysfunction is both a component of organ failure and a marker of systemic deterioration. The fact that serum creatinine change and urine-output dynamics also appeared among important predictors strengthens the interpretation that renal physiology is a key pre-onset signal.

Urine-output dynamics were especially informative. The range of 6-hour urine output over the previous 12 hours ranked third globally and was one of the top numeric dependence features. Low recent urine-output range increased predicted risk, while high variability had lower SHAP contribution. This is consistent with the role of urine output in SOFA renal scoring and with prior studies identifying urine output as important in sepsis prognosis (Hu *et al.* 2022; Zhang *et al.* 2025). Clinically, persistently low or nonrecovering urine output can reflect renal hypoperfusion, shock physiology, fluid balance abnormalities, or early acute kidney injury. A dynamic urine feature is therefore more informative than a single static urine value because it captures trajectory.

The urine-output features also required explicit outlier scrutiny. The raw landmark table contained implausibly high urine-derived values, with 6-hour urine output maxima above physiologic ranges and 0.7% of 12-hour urine-range landmarks exceeding 5,000 mL. Such values may reflect charting artifacts, duplicate aggregation, GU irrigation documentation, or non-urine output events mapped into structured output streams. For this reason, dependence plots were clipped to the 1st–99th percentile and a winsorized sensitivity analysis was added. Clipping urine-related features at training-set 1st–99th percentile limits produced h=6 AUROC 0.774 versus 0.772 for the no-year primary model, indicating that the main conclusion is not driven by extreme urine-output values. Future external implementations should apply site-specific urine-output plausibility filters before deployment.

The overlap between predictors and the Sepsis-3 label requires careful interpretation. Several important variables, including GCS, platelet count, bilirubin, creatinine, urine output, oxygenation-related proxies, and blood pressure, are direct SOFA components or close physiologic relatives of SOFA components. The model is therefore best described as forecasting future EHR-defined Sepsis-3 onset, partly by detecting early movement toward an organ-dysfunction threshold. This is clinically useful for early warning, but it is not the same as identifying a latent biological sepsis state independent of the label construction. The no-direct-SOFA-component sensitivity model reduced 6-hour AUROC to 0.758, indicating that SOFA-related physiology contributes meaningfully but does not solely explain the predictive signal.

Body temperature was the second most important global feature and had a nonlinear dependence pattern. Normothermic values around 36.5–37.0 °C were generally protective, while higher values increased risk. This is consistent with infection physiology and with Zhang *et al.* (2025), who identified temperature among the top features for multiple sepsis prognosis prediction in MIMIC-IV. The temperature feature was also technically strengthened by harmonizing Fahrenheit and Celsius charting; without that correction, temperature observation would have been artificially sparse and less reliable. The dependence curve should still be interpreted cautiously, because temperature can be affected by antipyretics, external warming/cooling, perioperative state, and measurement method.

Inflammatory and perfusion-related variables also behaved as expected. White blood cell count was ranked seventh globally and was a major positive contributor in the local SHAP example. Lactate itself was not in the global top 10, but lactate range over 12 hours and local lactate value contributed meaningfully. This is a useful distinction: lactate is often measured when clinicians are already worried, which is why measurement-process variables around lactate were excluded from CLEAN. The retained lactate value and lactate dynamics are physiologic variables; when present, they still provide useful signal. In the local case, lactate 3.5 mmol/L, WBC 20.2 K/uL, systolic blood pressure 88 mmHg, and GCS 3 formed a high-risk physiologic pattern.

Admission type remained influential in the primary no-year SHAP analysis, while deidentified admission-year group appeared only in the year-including internal benchmark. These variables should not be interpreted as direct causal mechanisms. Hospital admission type likely captures case-mix, route of presentation, acuity, and workflow differences between emergency, urgent, elective, observation, and surgical admissions. Anchor year group may capture secular changes in charting, coding, ICU practice, patient mix, or COVID-era effects; for that reason, it was removed from the primary deployment-oriented model and retained only as an internal benchmark. Prior MIMIC-IV fairness work has shown that demographic and administrative variables can reveal performance and fairness concerns in clinical models (Meng *et al.* 2022). For that reason, admission route and other administrative variables should be examined in subgroup calibration and fairness analyses before any clinical use.

The primary no-year model and temporal validation analyses address this concern directly. The year-including internal benchmark reached 6-hour AUROC 0.779, while removing deidentified admission-year group produced AUROC 0.772. Temporal validation, which trained on earlier deidentified year groups and tested on 2020–2022 without year group as a feature, produced 6-hour AUROC 0.762 and AUPRC 0.047. This temporal drop is expected under calendar-period shift and reinforces that the present model should be presented as a single-database benchmark requiring external validation, not as a transportable deployment model.

Local explanation and patient-level plausibility

The selected local SHAP case illustrates why patient-level explanations are useful. At ICU hour 8, the no-year model estimated a 45.2% probability of Sepsis-3 onset within the next 6 hours, and onset occurred 4.35 hours later. The largest positive driver was GCS 3, followed by hemoglobin, fever-range temperature, lactate, lactate range, emergency admission type, leukocytosis, low urine-output dynamics, hyperglycemia, bilirubin, tachycardia, and respiratory-rate features. This explanation does not merely list features; it describes a recognizable clinical syndrome of neurologic depression, inflammatory response, hypoperfusion, and early organ dysfunction.

The local explanation also reveals the difference between global and individual importance. Some variables that were not dominant globally, such as hemoglobin, glucose, bilirubin, or systolic blood pressure, contributed materially in this patient. These local attributions are context-dependent and can reflect interactions or correlated severity markers rather than isolated causal effects. This supports the use of both global and local SHAP in the manuscript. Global SHAP answers which variables matter on average; local SHAP answers why a specific patient was flagged at a specific time. In high-stakes early-warning settings, both views are necessary because clinicians need population-level trust and case-level interpretability.

Implications for clinical translation

The alarm-policy results should be interpreted as operating-point analysis, not as proof of clinical utility. The top-2% h=6 no-year policy gives a transparent trade-off: 17.8% horizon-concordant detection, 29.3% 0–24 hour detection, and 40.3% any-pre-onset enrichment at 15.5 alerts per 100 patient-days. It also requires about 10.5 emitted alerts per horizon-concordant detected case, which is a substantial operational cost. Prospective studies are needed to determine whether such alerts change clinician behavior, shorten time to antibiotics or cultures, reduce organ dysfunction, or improve mortality. The COMPOSER deployment study is an important reference point because it evaluated process and outcome changes after implementation rather than relying only on retrospective discrimination (Boussina *et al.* 2024).

The current model is intentionally simpler than recent multimodal or language-model systems (Shashikumar *et al.* 2025; Yang *et al.* 2026). A tabular LightGBM model can be trained reproducibly, explained with Tree SHAP, audited for leakage, and integrated into a landmark-alarm analysis with relatively low computational overhead. More complex models may improve discrimination by using notes, semantic abstraction, waveforms, or multimodal context, but they also introduce additional challenges in interpretability, maintenance, data availability, and external validation. The present work can therefore serve as a transparent baseline before moving to richer models.

Limitations

Several limitations should be acknowledged. First, this is a single-database retrospective study using MIMIC-IV v3.1. Although MIMIC-IV is widely used and carefully deidentified, it reflects one health system and its documentation practices. The added temporal validation reduces, but does not eliminate, concerns about transportability. External validation on other ICU databases is required before generalizing the operating thresholds. Second, Sepsis-3 onset labels are EHR-derived and depend on antibiotic, culture, and SOFA reconstruction rules. Misclassification is possible when infection suspicion is not captured by available medication or microbiology records, when cultures are delayed, or when organ dysfunction is incompletely observed.

Third, the forecastable cohort excludes Sepsis-3 onset within the first 4 ICU hours; therefore, the model is not intended to detect sepsis already present at ICU arrival. Fourth, death or discharge before a horizon is treated as censoring for the primary binary sepsis-onset task. This avoids labeling unobserved futures as non-sepsis, but death without observed sepsis can also be viewed as a competing event. A 6-hour composite sepsis-or-death sensitivity model achieved AUROC 0.808 and AUPRC 0.140, suggesting that broader deterioration is more discriminable than sepsis onset alone, but it is a different clinical endpoint. Fifth, the model uses structured data only. Clinical notes, imaging reports, waveform data, bedside nursing notes, and microbiology text could contain additional early information. Recent LLM-based and multimodal studies suggest that unstructured context can improve sepsis prediction in uncertain cases (Shashikumar *et al.* 2025; Yang *et al.* 2026).

Sixth, the CLEAN feature set deliberately removes measurement-process channels and vasopressor exposure to reduce leakage, but this conservative choice may also remove clinically useful real-time information. In deployment, one could evaluate multiple policies: a physiology-only warning model for early anticipation and a workflow-aware model for situational awareness. Seventh, SHAP explanations describe model behavior rather than causal effects. A high SHAP value for BUN, temperature, admission type, or urine-output range does not prove that changing that variable will change the patient's risk. The explanations should be interpreted as attribution within the fitted model. Eighth, alarm-policy performance was evaluated retrospectively. The false-alarm rate, clinician response, and patient benefit may change when alerts are displayed in real time. A prospective silent trial followed by pragmatic implementation would be required before clinical claims.

Future works

Future work should proceed in four directions. First, external validation should test the CLEAN model and alarm thresholds on independent ICU cohorts, ideally including eICU, AmsterdamUMCdb, or other contemporary ICU datasets. Second, subgroup performance should be evaluated by sex, age group, admission type, anchor year group, ICU type, and baseline severity to assess calibration and fairness. Third, trajectory-aware models such as recurrent neural networks, temporal convolutional networks, transformers, or semantic-abstraction LLM pipelines should be compared against the CLEAN LightGBM baseline under the same leakage-audited landmark labels. Fourth, a silent prospective evaluation should measure alert frequency, lead time, clinician agreement, and whether alerts would trigger clinically appropriate review before any interventional deployment.

Overall, the study supports a pragmatic conclusion: a carefully engineered, leakage-audited LightGBM model can provide interpretable Sepsis-3 onset forecasts from routine ICU physiology. Its discrimination is not as high as mortality prediction models, but that is expected because the task is earlier, rarer, and more clinically prospective. The strongest scientific contribution is the combination of dynamic multi-horizon labels, CLEAN feature auditing, alarm-policy evaluation, and SHAP-based physiologic interpretation. This combination makes the model a transparent benchmark and a foundation for future externally validated early-warning research.

CONCLUSION

This study developed and evaluated a leakage-audited, interpretable machine-learning pipeline for forecasting future Sepsis-3 onset in adult ICU patients using MIMIC-IV v3.1. Rather than treating sepsis prediction as a single static classification task, the analysis formulated prediction as a repeated landmark problem: at each ICU hour, the model estimated whether Sepsis-3 onset would occur within clinically relevant future windows of 4, 6, 12, and 24 hours. This design allowed the model to represent the dynamic nature of critical illness while preserving patient-level separation between training, validation, and test cohorts.

The main empirical finding is that routinely collected ICU physiology contains signal for future EHR-defined Sepsis-3 onset. The no-year CLEAN LightGBM model achieved test AUROCs of 0.782, 0.772, 0.750, and 0.720 across the 4-, 6-, 12-, and 24-hour horizons, respectively. The 6-hour horizon provided the most clinically balanced operating point, but alarm interpretation depends on the detection window. Under a top-2% alarm policy with a 6-hour cooldown, the model detected 17.8% of sepsis cases within the horizon-concordant 0–6 hour window, 29.3% within 0–24 hours, and 40.3% under a broader any-pre-onset definition, with an alarm burden of 15.5 alerts per 100 patient-days.

A central contribution of the study is the explicit separation of physiologic prediction from workflow-driven leakage. Removing vasopressor exposure, measurement-recency variables, and observation-mask channels caused only minimal performance degradation, with AUROC decreases below 0.01 across all horizons. This result suggests that model performance was not primarily driven by clinician test-ordering behavior or treatment proxies. For academic and clinical interpretation, this is important because it strengthens the claim that the model learned from patient state and physiologic trajectories rather than from retrospective artifacts of care delivery.

The explainability analyses further support physiologic face validity. Global SHAP analysis identified neurologic status, body temperature, urine-output dynamics, blood urea nitrogen, inflammatory markers, vital signs, and acid-base or lactate-related trends as major contributors. These variables align with known domains of Sepsis-3 organ dysfunction and with prior interpretable sepsis studies. Local SHAP analysis in an example patient showed that the model's high-risk prediction was driven by depressed consciousness, fever-range temperature, elevated lactate, leukocytosis, borderline hypotension, abnormal renal-output dynamics, and emergency admission context. Dependence plots showed nonlinear effects for the most influential numeric predictors.

Taken together, these findings support the proposed pipeline as an internally and temporally validated benchmark for Sepsis-3 onset forecasting. The model is not presented as a clinical decision-support system; it remains a retrospective, single-database study requiring external validation, prospective silent evaluation, subgroup calibration monitoring, and workflow testing. Nevertheless, the combination of dynamic horizon-specific labels, conservative feature auditing, interpretable LightGBM modeling, SHAP-based explanation, and alarm-policy evaluation provides a reproducible framework for future early-warning research. The results indicate that a physiology-based model can identify a subset of ICU patients at increased risk of future EHR-defined Sepsis-3 onset while producing explanations that can be inspected at both cohort and patient levels.

Supplementary Material

The submission package includes supplementary CSV and Markdown files containing baseline characteristics, patient-level split counts, label observability by split and horizon, the patient-level bootstrap table, decile calibration, temporal-validation results with bootstrap intervals, temporal alarm-policy results, clinical-score and calibrated logistic-regression baselines, sensitivity analyses, revised alarm-window metrics with bootstrap intervals, subgroup performance and subgroup alarm summaries, sepsis-onset timing detection, no-year SHAP summaries, categorical SHAP summaries, urine-output plausibility checks, competing-risk sensitivity results, a related-work comparison table, a CLEAN feature dictionary with source tables and MIMIC item identifiers, package versions, a TRIPOD-AI/CLAIM-oriented reporting checklist, a synthetic landmark-labeling smoke-test example, a reference-metadata spot-check note, and a reproducibility note documenting landmark generation, labeling, model settings, class-imbalance handling, and censoring rules.

Availability of data and material

This study uses MIMIC-IV v3.1, which is available through PhysioNet to credentialed users who complete the required data-use training and data-use agreement. Code for cohort construction, feature generation, landmark labeling, model training, evaluation, and figure generation will be released in a public repository upon acceptance. An anonymized code archive will be made available to reviewers as confidential supplementary material where permitted by journal policy. The repository will not contain MIMIC-IV data and must be run by credentialed users with local MIMIC-IV access. The supplementary material lists the derived feature dictionary, item identifiers, package versions, random seed, model hyperparameters, labeling rules, and a synthetic smoke-test example needed to verify the labeling logic without access to protected clinical data.

Acknowledgements

The authors would like to thank researcher who publicly share their dataset in kaggle. The authors acknowledge Sakarya University of Applied Sciences (https://subu.edu.tr/) for the technical support provided to publish the present manuscript.

Funding

This research received no external funding.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical standard

The authors have no relevant financial or non-financial interests to disclose.

LITERATURE CITED

Boussina, A., S. P. Shashikumar, A. Malhotra, R. L. Owens, R. El-Kareh, *et al.*, 2024 Impact of a deep learning sepsis prediction model on quality of care and survival. npj Digital Medicine **7**: 14.

Duval, L., A. Villié, F. Zheng, G. Terraz, S. Blein, *et al.*, 2025 Early prediction of vasopressor initiation in ICU sepsis patients using an interpretable EHR-based ML model. BMC Medical Informatics and Decision Making **25**: 442.

Gao, J., Y. Lu, N. Ashrafi, I. Domingo, K. Alaei, *et al.*, 2024a Prediction of sepsis mortality in ICU patients using machine learning methods. BMC Medical Informatics and Decision Making **24**: 228.

Gao, Y., C. Wang, J. Shen, Z. Wang, Y. Liu, *et al.*, 2024b Systematic review and network meta-analysis of machine learning algorithms in sepsis prediction. Expert Systems with Applications **245**: 122982.

Hu, C., L. Li, W. Huang, T. Wu, Q. Xu, *et al.*, 2022 Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. Infectious Diseases and Therapy **11**: 1117–1132.

Islam, K. R., J. Prithula, J. Kumar, T. L. Tan, M. B. I. Reaz, *et al.*, 2023 Machine learning-based early prediction of sepsis using electronic health records: a systematic review. Journal of Clinical Medicine **12**: 5658.

Johnson, A., L. Bulgarelli, T. Pollard, B. Gow, B. Moody, *et al.*, 2024 MIMIC-IV. Version 3.1.

Johnson, A. E. W., L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, *et al.*, 2023 MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data **10**: 1.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, *et al.*, 2017 LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154.

Liu, Z., W. Shu, T. Li, X. Zhang, and W. Chong, 2025 Interpretable machine learning for predicting sepsis risk in emergency triage patients. Scientific Reports **15**: 887.

Lu, X., H. Kang, D. Zhou, and Q. Li, 2022 Prediction and risk assessment of sepsis-associated encephalopathy in ICU based on interpretable machine learning. Scientific Reports **12**: 22621.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, *et al.*, 2020 From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence **2**: 56–67.

Meng, C., L. Trinh, N. Xu, J. Enouen, and Y. Liu, 2022 Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Scientific Reports **12**: 7166.

Seymour, C. W., V. X. Liu, T. J. Iwashyna, F. M. Brunkhorst, T. D. Rea, *et al.*, 2016 Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA **315**: 762–774.

Shan, W., D. Sun, and Z.-P. Liu, 2024 Predicting sepsis onset in ICU patients using machine learning and feature section: a case study of MIMIC-IV data. In *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pp. 546–551.

Shashikumar, S. P., S. Mohammadi, R. Krishnamoorthy, A. Patel, G. Wardi, *et al.*, 2025 Development and prospective implementation of a large language model based system for early sepsis prediction. npj Digital Medicine **8**: 290.

Singer, M., C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, *et al.*, 2016 The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA **315**: 801–810.

Yang, Y., B. Yi, and Y. Chen, 2026 Lightweight large language models for early sepsis prediction via a semantic abstraction rule engine. Scientific Reports .

Zhang, S.-Z., H.-Y. Ding, Y.-M. Shen, B. Shao, Y.-Y. Gu, *et al.*, 2025 Harness machine learning for multiple prognoses prediction in sepsis patients: evidence from the MIMIC-IV database. BMC Medical Informatics and Decision Making **25**: 152.

Zhang, Y., T. Kirchler, and A. P. Wang, 2026 Evaluating artificial intelligence for sepsis prediction in emergency departments: a systematic review and meta analysis. Journal of Medical Systems **50**: 49.

Zubair, M., I. Din, N. Sarwar, B. Elov, and Z. Trabelsi, 2025 Revolutionizing sepsis diagnosis using machine learning and deep learning models: a systematic literature review. BMC Infectious Diseases **25**: 1396.

How to cite this article: Mansour, M., and Donmez T. B. An Explainable Leakage-Audited Framework for Multi-Horizon Sepsis-3 Onset Forecasting from Complex ICU Dynamics. *Chaos and Fractals*, 3(2), 76-91, 2026.