

Development and validation of a simplified two-point sampling strategy for predicting in-hospital mortality in elderly ICU sepsis patients: a multicenter retrospective cohort study

Weixi Guo

Central South University

Jian Yang

Central South University

Yilin Xia

Central South University

Bei Li

libeiwu@126.com

Central South University

Research Article

Keywords: Sepsis, Machine learning, Temporal sampling density, Elderly patients, Intensive care unit, Interpretability

Posted Date: June 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9697683/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on August 19th, 2026. See the published version at <https://doi.org/10.1186/s12911-026-03767-9>.

2 **Development and validation of a simplified two-point sampling**
3 **strategy for predicting in-hospital mortality in elderly ICU sepsis**
4 **patients: a multicenter retrospective cohort study**

5 Weixi Guo¹, Jian Yang¹, Yilin Xia¹, Bei Li^{1*}

6 ¹ Department of Biomedical Informatics, School of Life Sciences, Central South University,
7 No.172 Tongzipo Road, Changsha 410000, China

8 ***Corresponding author:**Bei Li, Department of Biomedical Informatics, School of Life Sciences,
9 Central South University, No.172 Tongzipo Road, Changsha 410000, China

10 E-mail: libeiwu@126.com Phone and fax:+86 18874811152

11 **Abstract**

12 **Background:** Sepsis is a leading cause of death in the intensive care unit (ICU), with
13 significantly higher rates in elderly patients due to immunosenescence and greater
14 comorbidity burden. Existing machine learning models for ICU mortality prediction
15 predominantly rely on fixed-time-window feature extraction without evaluating how
16 temporal sampling density influences performance. This study aimed to compare the effects
17 of different temporal sampling density strategies on in-hospital mortality prediction in
18 elderly sepsis patients and identify the critical sampling density at which information
19 saturates.

20 **Methods:** This retrospective cohort study used the Medical Information Mart for Intensive
21 Care IV database as the development set and the Electronic ICU Collaborative Research
22 Database as the external validation set. Patients aged 65 years or older who met the Sepsis-
23 3 criteria and had an ICU stay of at least 24 hours were included. A fixed-architecture long
24 short-term memory network was used to compare six temporal sampling strategies (1-
25 point, 2-point, 3-point, 4-point, 6-point, and 24-point) under a unified feature set and

training strategy. Model performance was evaluated in terms of the area under the receiver operating characteristic curve, the area under the curve, the F1 score, and the Brier score. Interpretability analysis was performed to identify the core drivers of the optimal model.

Results: Of the 546028 screened records, 20163 elderly sepsis patients were included. The 2-point group (admission and 24-hour) achieved optimal discrimination (area under the receiver operating characteristic curve 0.814, 95% confidence interval 0.809–0.818), with the largest improvement over the single-point group (0.122; $P < 0.001$). The difference versus the 24-point hourly group was not significant ($P = 0.961$), indicating predictive equivalence. Adding intermediate sampling points beyond 2 points yielded no additional gain. Sensitivity analysis confirmed that the 24-hour endpoint contains richer prognostic information than earlier timepoints do. External validation ($n = 11039$) confirmed these findings. The 24-hour Glasgow Coma Scale score, blood urea nitrogen, and lactate level were the core predictive indicators.

Conclusions: In elderly ICU sepsis patients, a simplified two-point sampling strategy (admission baseline and 24-hour endpoint) achieves predictive equivalence to hourly dense monitoring. Intermediate sampling points provide no additional gain, suggesting that simplified sampling is sufficient under ICU data sparsity constraints.

Keywords: Sepsis; Machine learning; Temporal sampling density; Elderly patients; Intensive care unit; Interpretability

Background

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection and is one of the leading causes of death among patients in intensive care units (ICUs) worldwide [1]. An estimated 48.9 million new sepsis cases occur globally each year, resulting in approximately 11 million deaths that account for nearly 20% of all-cause mortality [2]. Elderly populations exhibit significantly greater sepsis incidence and mortality than the general population owing to immunosenescence, diminished physiological reserve, and greater comorbidity burden [3,4]. Studies have reported in-hospital mortality rates of 30–50% among

sepsis patients older than 65 years, with mortality increasing progressively with advancing age [5]. Therefore, early identification of elderly sepsis patients at high risk of death is clinically important for optimizing resource allocation and improving outcomes.

Currently, conventional severity scoring tools such as the Acute Physiology and Chronic Health Evaluation II (APACHE II) and the Sequential Organ Failure Assessment (SOFA) are widely used to assess mortality risk in ICU patients [6,7]. However, these traditional scoring systems have significant limitations: their scoring variables are relatively fixed, making it difficult to adequately capture dynamic disease evolution; they exhibit suboptimal predictive accuracy in specific subpopulations, such as elderly sepsis patients; and they lack the capacity to leverage temporal data [8,9]. With the increasing availability of electronic health record (EHR) data and advances in artificial intelligence, machine learning (ML) methods have demonstrated substantial potential for ICU prognostic prediction [10], and the predictive performance of automatically extracted EHR data has been shown to be comparable to that of manually curated data [46].

In recent years, numerous studies have employed ML models to predict mortality risk in intensive care unit (ICU) sepsis patients, using diverse algorithmic frameworks ranging from traditional machine learning (e.g., XGBoost, random forest) to deep learning (e.g., long short-term memory [LSTM], gated recurrent units [GRUs], and transformers) [11–15]. A systematic review encompassing 31 studies and more than 1.47 million sepsis patients reported that ML models achieved AUROC values ranging from 0.62--0.90 for sepsis-related mortality prediction, generally outperforming conventional scoring systems (e.g., SOFA AUROC 0.47--0.86) [16].

The application of interpretability analysis methods such as SHAP has further enhanced the clinical credibility and utility of these models [17,18]. Nevertheless, existing studies have important shortcomings: first, only 29% of studies have completed external validation, and 61% present a high risk of bias, raising concerns about model generalizability [16,19]; second, most models employ fixed-time window "snapshot" feature extraction (e.g., within 24 hours of ICU

77 admission) without adequately considering how temporal sampling density influences model
78 performance [20,21].

79 In clinical practice, the acquisition frequency of physiological indicators varies
80 substantially—from continuous real-time monitoring to laboratory tests performed only
81 every few hours. Although efficient prognostic prediction is limited, the use of limited
82 clinical data remains a critical clinical challenge [22,23]. Although prior studies have
83 explored the impact of different time window sizes on sepsis prediction performance
84 [23,24], few have systematically investigated whether prediction performance exhibits a
85 saturation effect across the complete gradient from simplified (single-point) to dense
86 (hourly) sampling within a fixed observation window. Notably, the transition from an
87 admission static baseline (t_0) to incorporating follow-up dynamic information has already
88 produced a significant improvement in predictive performance [8,11]; however, the
89 marginal benefit of adding intermediate sampling timepoints after obtaining baseline and
90 endpoint information remains unclear. In fact, ICU temporal data contain two
91 fundamentally different acquisition modes: "streaming" vital signs continuously recorded
92 by monitors (missing rate <6%) and "pulsed" laboratory indices (e.g., lactate, creatinine,
93 international normalized ratio [INR]) typically collected only 1–2 times daily, with
94 instantaneous missing rates at any single-hour slice reaching 84–95% and Glasgow Coma
95 Scale (GCS) missing rates of 61–75% (Supplementary Material 2). Consequently, overly
96 dense sampling schemes inevitably rely heavily on last observation carried forward (LOCF)
97 imputation, which may degrade model performance by amplifying noise [25,26]. Therefore,
98 determining the critical sampling density at which information saturates for in-hospital
99 mortality prediction in elderly sepsis patients is essential for constructing simplified,
100 efficient, and clinically deployable prediction models.

This study aimed to use two large publicly available ICU databases, the MIMIC-IV and eICU-CRD, to compare the effects of six temporal sampling density strategies (1-point, 2-point, 3-point, 4-point, 6-point, and 24-point) on in-hospital mortality prediction performance in elderly sepsis patients via a fixed-architecture LSTM network. We hypothesized that, within a fixed 24-hour observation window, a simplified two-point sampling strategy (admission baseline plus 24-hour treatment endpoint) would achieve predictive equivalence to hourly dense monitoring and that further increasing intermediate sampling points would not provide additional predictive gain. Through dual-database validation and SHAP interpretability analysis, this study provides an efficient data acquisition approach for the clinical deployment of ICU sepsis mortality prediction models.

Methods

Study design and study population

This study was designed and reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis—Artificial Intelligence (TRIPOD+AI) 2024 update [27] and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [28]. Within this framework, this retrospective cohort study used data from two publicly available critical care databases: the Medical Information Mart for Intensive Care IV (MIMIC-IV, version 3.1) and the Electronic ICU Collaborative Research Database (eICU-CRD, version 2.0). The MIMIC-IV served as the model development set, and the eICU-CRD was used for external validation. This study aimed to explore the effects of different temporal sampling densities on in-hospital mortality prediction performance in elderly sepsis patients. To eliminate interference from the model algorithm itself, a fixed-architecture deep learning model was used as a temporal feature detector to systematically compare the predictive performance of six temporal sampling densities (1-point, 2-point, 3-point, 4-point, 6-point, and 24-point). All the models were trained and evaluated under a unified feature set, model

architecture, and training strategy, with temporal sampling density as the sole variable. The overall technical workflow is illustrated in Figure 1 and encompasses six key stages: cohort construction, temporal sampling design, model development, internal validation, interpretability analysis, and clinical utility assessment.

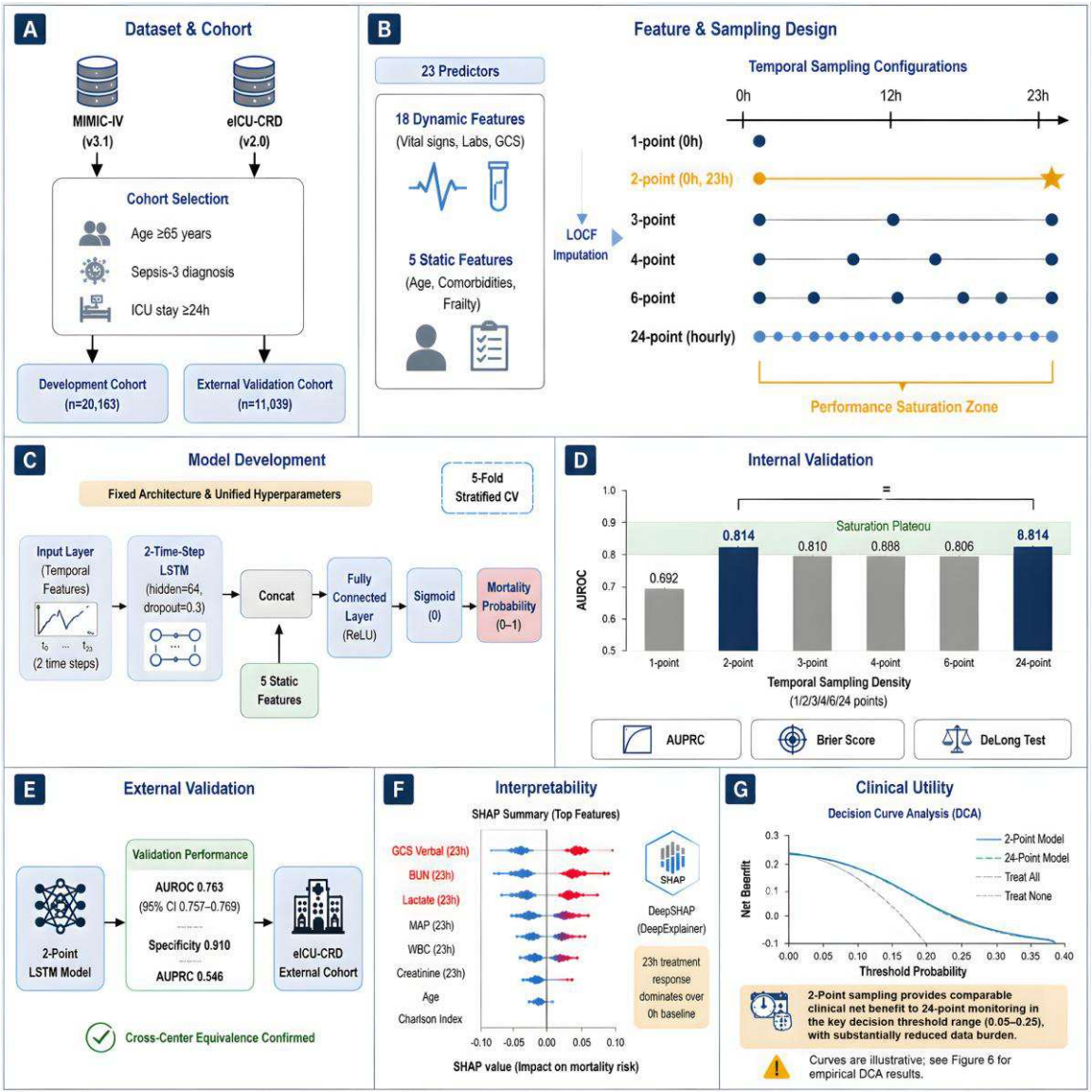


Figure 1 Study workflow. A total of 20163 elderly sepsis patients were retrospectively enrolled from MIMIC-IV v3.1 as the development cohort, and 11039 patients from eICU-CRD v2.0 were used as the external validation cohort. Twenty-three clinical variables were extracted at six candidate temporal sampling densities within the first 24 hours after ICU admission. The last

observation carried forward (LOCF) method was used to handle missing data. A fixed-architecture unidirectional long short-term memory (LSTM) network with two hidden layers (64 units each) and a fused input vector of 87 dimensions was employed for in-hospital mortality prediction. Model performance was evaluated via the area under the receiver operating characteristic curve (AUROC) with 95% confidence intervals derived from 1,000 bootstrap resamples. DeLong's test with Bonferroni correction (significance threshold $\alpha=0.0033$) was applied for pairwise comparisons between different sampling strategies. SHapley Additive exPlanations (SHAP) analysis and decision curve analysis (DCA) were performed to interpret model predictions and assess their clinical utility.

The MIMIC-IV database comprises the clinical records of patients admitted to the intensive care units of Beth Israel Deaconess Medical Center (Boston, USA) from 2008--2019, containing more than 500,000 ICU admissions. The eICU-CRD database is a multicenter telemedicine database containing clinical information for more than 200,000 patients from 335 ICUs across 208 hospitals in the United States from 2014–2015. Both databases have received approval from their respective institutional review boards, have been deidentified, and have waived informed consent. All data access was completed through PhysioNet credentialing.

The inclusion criteria were as follows: age ≥ 65 years at ICU admission; meeting the Sepsis-3 diagnostic criteria (infection resulting in an increase in the Sequential Organ Failure Assessment [SOFA] score of ≥ 2 points); ICU stay ≥ 24 hours; complete vital sign records at 0 hours after admission; and at least one recorded timepoint between 12 and 23 hours. The exclusion criteria were missing in-hospital mortality outcome labels, missing key static features (age or comorbidity information), physiologically implausible values, and data coverage below 50% during the 0- to 24-hour period after admission.

Data collection and preprocessing

Data, including demographic characteristics, vital signs, laboratory results, and comorbidity information, were extracted from the EHR systems of the MIMIC-IV and eICU-CRD databases within 24 hours of ICU admission. A total of 23 predictors were included according to three principles: (i) established core prognostic factors from the sepsis mortality literature, including organ dysfunction indicators involved in the Sepsis-3 diagnostic criteria [1], physiological and laboratory parameters covered by the SOFA scoring system [7], acute physiological variables in the APACHE II framework [6], tissue perfusion markers such as lactate [29], and neurological function assessment indicators such as the Glasgow Coma Scale (GCS) [30]; (ii) clinical routine accessibility, meaning that all indicators are routinely collected at the bedside within 24 hours of ICU admission; and (iii) collinearity exclusion, whereby candidate variables underwent Pearson correlation analysis and highly collinear variables with $|r| > 0.8$ or variance inflation factor (VIF) > 10 were excluded (e.g., total GCS score, activated partial thromboplastin time). The complete candidate variable pool and selection rationale are detailed in Supplementary Material 1. The final set of 23 predictors comprised 18 dynamic features and 5 static features. Dynamic features included heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, respiratory rate, pulse oxygen saturation (SpO_2), body temperature, oxygenation index (PaO_2/FiO_2), lactate, serum creatinine, blood urea nitrogen (BUN), white blood cell count, hemoglobin, platelet count, international normalized ratio (INR), GCS eye opening response, GCS verbal response, and GCS motor response. Static features included age, diabetes history, heart failure history, Charlson Comorbidity Index [31], and frailty status [32].

Data processing comprises multiple steps. For dynamic features, missing values were imputed via the last observation carried forward (LOCF) method; that is, the most recent valid observation for each patient was used to fill missing data, and any remaining missing values after LOCF imputation were filled via the median of the training set from 5-fold cross-validation. Missing static feature values were imputed using the training set mode (for binary variables) or median

(for continuous variables). For continuous variables, winsorization was applied via the 1st and 99th percentiles of the training set to manage outliers, with extreme values beyond this range replaced by the corresponding percentile values. All the features were z score standardized on the training set, and the standardization parameters were directly applied to the external validation set to ensure independence of the validation process and prevent data leakage.

Outcome definition

The primary outcome of this study was all-cause in-hospital mortality during ICU hospitalization, which was defined as death from any cause while the patient was in the ICU. The mortality label was extracted directly from the hospital discharge status field in the database, with survival coded as 0 and death coded as 1. For the MIMIC-IV database, in-hospital mortality status was determined via `hospital_expire_flag`; for the eICU-CRD database, it was determined via the actual hospital mortality field. All outcome labels were extracted before feature engineering to ensure label independence.

Temporal sampling density experimental design

The core experimental design of this study aimed to isolate the independent effect of temporal sampling density on model predictive performance. To achieve this goal, a unified fixed-architecture LSTM model was employed, with completely consistent model hyperparameters, feature sets, training strategies, and evaluation procedures maintained across all experimental groups; the only varying parameter was the number of temporal sampling points in the input time series data. On the basis of actual ICU clinical workflows and data acquisition frequencies, six progressive temporal sampling schemes were designed: a 1-point group (admission baseline only, 0 h); a 2-point group (0 h and 23 h); a 3-point group (0 h, 12 h, and 23 h); a 4-point group (0 h, 8 h, 16 h, and 23 h); a 6-point group (0 h, 5 h, 10 h, 14 h, 19 h, and 23 h); and a 24-point group (hourly discrete sampling from 0--23 h). The 0-hour timepoint corresponds to the admission baseline assessment, and the 23-hour timepoint corresponds to the routine re-evaluation at the end of the first treatment day after ICU admission.

Data extraction for each temporal sampling group followed a uniform LOCF forward-fill rule: at each prespecified sampling timepoint, the current or most recent prior valid observation was extracted; if no record existed at that timepoint or at any preceding timepoint, the training set median was used for imputation. This mechanism ensured that differences between sampling density groups arose solely from the sampling frequency rather than from the imputation strategy. The six temporal sampling schemes, including their respective timepoints, are detailed in Table 1.

Table 1 Temporal sampling strategies of the six experimental groups

Group	Number of points	Sampling timepoints (h)	Number of timesteps
1-point	1	0	1
2-point	2	0, 23	2
3-point	3	0, 12, 23	3
4-point	4	0, 8, 16, 23	4
6-point	6	0, 5, 10, 14, 19, 23	6
24-point	24	Hourly (0–23)	24

The 0-hour timepoint corresponds to the admission baseline; the 23-hour timepoint corresponds to the end of the first ICU treatment day.

ML model development and evaluation

This study employed a unidirectional long short-term memory (LSTM) network [33] as the temporal sequence feature extractor. The LSTM model input comprises dynamic feature matrices arranged in chronological order (number of timepoints $\times$ 18 dimensions). Static features (5 dimensions) were concatenated with the final LSTM hidden state (64 dimensions) and the dynamic features at the last timestep (18 dimensions) after sequence processing, forming a fused vector (total dimension 87) that was fed into the classification head. The specific architecture

included two unidirectional LSTM layers with a hidden dimension of 64 and an interlayer dropout rate of 0.3 and a classification head consisting of a fully connected layer (with an input of 87 dimensions and an output of 32 dimensions), a ReLU activation function, a dropout layer (with a rate of 0.2), and an output layer (32 dimensions to 1 dimension), with the final death probability predicted via a sigmoid function.

Model development was conducted on the MIMIC-IV database. A 5-fold stratified cross-validation strategy was adopted to ensure that the proportion of death and survival samples in each fold was consistent with that of the full dataset. In each cross-validation iteration, 4 folds were used for model training, and the remaining fold served as the validation set for hyperparameter tuning and early stopping decisions; after completion of the 5-fold validation, the prediction results from all folds were pooled for performance evaluation. To eliminate randomness from data partitioning, a random seed of 42 was set to ensure reproducibility of the experimental results. Model training used the Adam optimizer with an initial learning rate of 1×10^{-3} , $\beta_1=0.9$, and $\beta_2=0.999$; the early stopping strategy was set to terminate training when the validation set AUROC showed no improvement for 5 consecutive epochs; the batch size was 64; and the maximum number of training epochs was 100. The model uses a binary cross-entropy loss function, and class weights are dynamically adjusted during training to partially mitigate the positive-negative sample imbalance.

Model performance was comprehensively evaluated via the following metrics: the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), the F1 score, and the Brier score [37]. AUROC and AUPRC reflect the comprehensive discrimination ability of the model; the F1 score reflects the balance between precision and recall; and the Brier score reflects the calibration of the predicted probabilities. All the performance metrics were calculated on the basis of the pooled prediction results from 5-fold cross-validation. The 95% confidence intervals (CIs) were constructed via 1000 bootstrap resamplings [34] to assess the stability of the model performance estimates. The statistical significance of AUROC

differences between groups was assessed via the DeLong test [35], with a two-sided significance level of $\alpha=0.05$. Model development, training, and evaluation were implemented via the PyTorch framework (version 1.12.0), and all the experiments were performed on a Linux workstation equipped with an NVIDIA GeForce RTX 3090 GPU. The complete experimental code and data preprocessing scripts have been made publicly available on GitHub (<https://github.com/wangjunnyworld/sepsis-temporal-sampling>) to ensure study reproducibility.

Benchmark model comparison

The benchmark model comparison framework of this study centered on temporal sampling density as the core variable. The six temporal sampling density schemes (1-point, 2-point, 3-point, 4-point, 6-point, and 24-point) constituted mutually controlled benchmark model groups, in which the 1-point group (admission baseline only) represented the traditional static prediction mode without dynamic follow-up information, the 24-point group (hourly continuous sampling) represented the idealized full-density dynamic monitoring mode, and the 2-point, 3-point, 4-point, and 6-point groups represented progressively denser sampling strategies. By systematically comparing the AUROC, AUPRC, F1 score, and Brier score across the six groups, the critical sampling density at which the prediction performance saturates was identified, and the full-spectrum performance gradient from simplified to dense sampling was quantified.

The statistical significance of performance differences between groups was assessed via the DeLong test for pairwise comparisons, with a total of 15 pairwise comparisons across the six groups; after Bonferroni correction [36], the significance threshold was $\alpha=0.0033$. For sampling strategies approaching performance saturation, 1000 bootstrap resamplings were further used to evaluate the 95% confidence interval of AUROC differences between adjacent sampling density groups to confirm performance equivalence or comparison results presented as forest plots, with AUROC values and their 95% confidence intervals

displayed side by side across sampling density groups to facilitate intuitive identification of the performance inflection point. All performance evaluations were independently repeated on the external validation set (eICU-CRD) to confirm the robustness of the main findings in cross-center scenarios.

Interpretability analysis of the optimal ML model

To reveal the decision mechanism of the optimal temporal sampling density model (2-point sampling group), this study employed the SHapley Additive exPlanations (SHAP) method for post hoc interpretability analysis of the model prediction results. SHAP is based on cooperative game theory principles and assigns an additive attribution value (SHAP value) to each feature in individual prediction instances by calculating the weighted average marginal contribution of each feature across all possible feature combinations, thereby quantifying the positive or negative driving effect of each feature on the model output. Given that the LSTM architecture used in this study is a deep learning model, the DeepSHAP algorithm (implemented via the DeepExplainer interface in the Python SHAP library) was used for model interpretation, with the background distribution set as a random subsample of 1000 samples from the training set.

The interpretability analysis comprised three levels. First, global feature importance ranking: the mean absolute SHAP value for each feature across all validation samples was calculated to obtain the feature importance ranking, identifying the core clinical variables driving death prediction. Second, single-feature dependence analysis was performed: key features ranked highly in global importance (e.g., GCS verbal score, BUN, lactate) were selected, and SHAP dependence plots were drawn to reveal the functional relationship between feature value and predicted risk, identifying patterns such as linear increase, stepwise transition, or threshold effects. The third step is clinical threshold mining: critical values at which the risk probability underwent significant turning points were identified within SHAP dependence plots, and the pathophysiological mechanisms underlying threshold effects were interpreted in conjunction with clinical physiological knowledge. Interpretability analyses were performed on the pooled results of 5-fold

cross-validation from the MIMIC-IV development set to ensure the robustness of feature attribution. Analysis was implemented via the Python SHAP library (version 0.41.0).

Sensitivity analysis

To verify the robustness of the performance advantage of the optimal two-point sampling scheme (0 h and 23 h) and to exclude the possibility that this result depended on a specific timepoint selection, multiple sensitivity analysis experiments were designed. With the 0-hour baseline timepoint fixed, the 6-hour, 12-hour, 18-hour, and 23-hour timepoints were used as the second timepoint to construct single-variable-variation 2-point sampling control groups, and the effects of different second-timepoint selections on the model AUROC were evaluated. In addition, to assess the sensitivity of model performance to randomness in data partitioning, 5-fold cross-validation experiments were repeated using different random seeds (10, 20, 30, 40, 50), and the variation range of the AUROC was observed. The above sensitivity analyses were all completed on the MIMIC-IV development set and simultaneously reproduced on the external validation set (eICU-CRD) to confirm the cross-center robustness of the main findings.

Statistical analysis

Continuous variables are presented as the means $\pm$ standard deviations or medians (interquartile ranges), and categorical variables are presented as frequencies (percentages). Between-group comparisons were performed via independent-samples t tests (normal distribution) or Mann–Whitney U tests (nonnormal distribution) for continuous variables and chi-square tests or Fisher's exact tests for categorical variables. The 95% confidence intervals for model performance metrics were constructed via 1000 bootstrap resamplings; in each resampling, samples equal in size to the original dataset were drawn with replacement, the target metric was calculated, and the 2.5th and 97.5th percentiles were used as the lower and upper bounds of the confidence interval. The statistical significance of AUROC differences between groups was assessed via the DeLong test; multiple comparison correction was performed via the Bonferroni method, with the corrected significance threshold set at $\alpha=0.05$ divided by the number of comparisons. All the statistical

analyses were implemented via the Python SciPy library (version 1.9.0) and the scikit-learn library (version 1.1.0), with two-sided tests and a significance level of $\alpha=0.05$.

Results

Study population

In the MIMIC-IV cohort, 546028 ICU hospitalization records were initially screened, 41,295 of which met the Sepsis-3 diagnostic criteria. After further screening for patients aged ≥ 65 years, 21,923 patients were identified; after confirming valid feature records within 24 hours, 21,921 patients entered the initial screening queue. Subsequent stepwise exclusion removed 129 patients without 0-hour baseline data, 487 patients lacking 12- to 23-hour follow-up data, 266 patients with 0- to 23-hour data coverage below 50%, and 876 patients with incomplete core vital signs at 0 hours. Ultimately, 20163 patients meeting all the inclusion criteria were included in the model development cohort, comprising 16393 survivors (81.3%) and 3770 nonsurvivors (18.7%). On the basis of the 23 predictors included in the final model, the events-per-variable (EPV) ratio was approximately 164:1, far exceeding the conventional recommended threshold for prediction model construction ($EPV \geq 10$), thereby ensuring the robustness of model estimation and effectively reducing overfitting risk. In the eICU-CRD cohort, after identical screening procedures, 11039 patients were ultimately included in the external validation cohort, comprising 8,106 survivors (73.4%) and 2,933 nonsurvivors (26.6%). The two groups were generally comparable in terms of age, sex distribution, and major clinical characteristics. Significant differences in the Sequential Organ Failure Assessment (SOFA) score, serum lactate level, blood urea nitrogen (BUN) level, and Glasgow Coma Scale (GCS) score (all $P < 0.001$) were detected between the death and survival groups, with patients in the death group exhibiting more severe organ dysfunction and metabolic disturbances. The baseline characteristics of the study population, stratified by survival status, are summarized in Table 2 (see end of manuscript). The patient screening flowchart is shown in Figure 2.

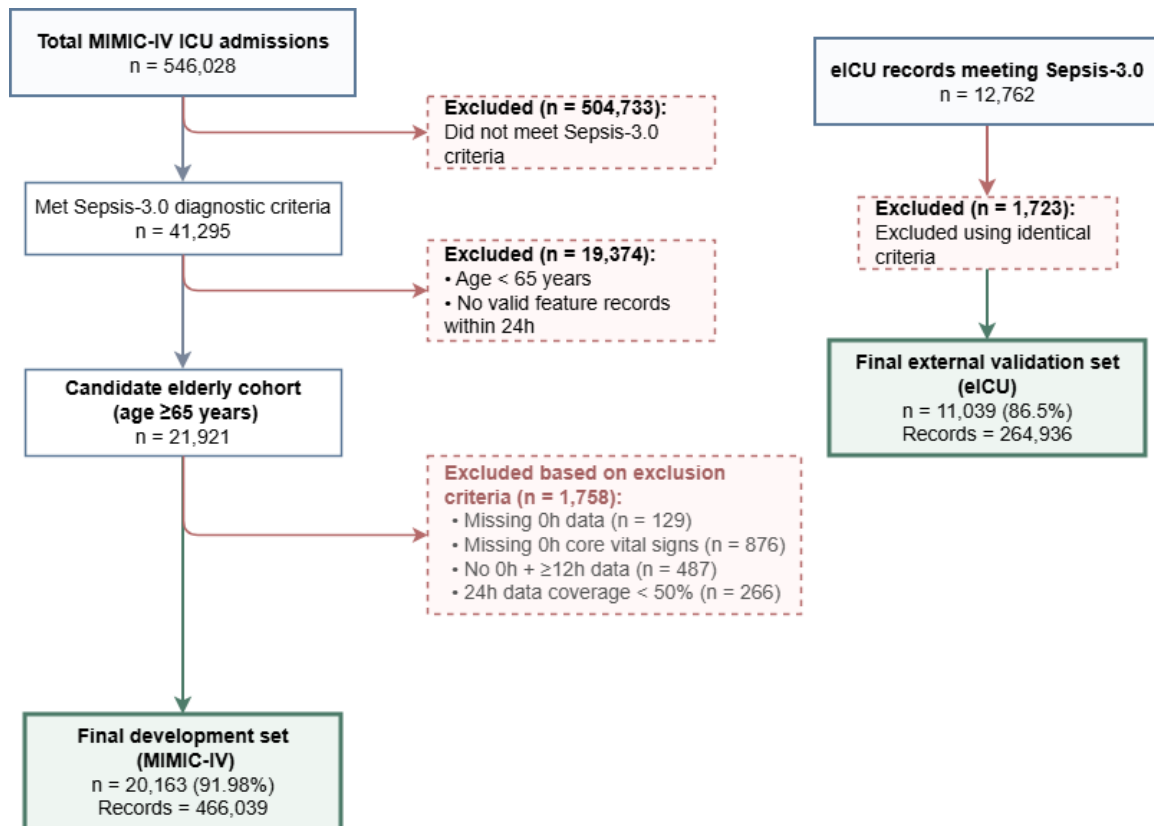


Figure 2 Patient screening flowchart for the multicenter temporal prediction cohort of elderly sepsis patients. Left panel: The MIMIC-IV database initially contains 546028 ICU hospitalization records; after multilayer screening (Sepsis-3 diagnostic criteria, age ≥ 65 years, key feature completeness, and data quality), 20163 patients were ultimately included in the model development set. Right panel: The eICU-CRD database initially contained 12762 records meeting the Sepsis-3 criteria; after identical screening, 11039 patients were included in the external validation set. Abbreviations: MIMIC-IV, Medical Information Mart for Intensive Care IV; eICU-CRD, Electronic ICU Collaborative Research Database; Sepsis-3, Third International Consensus Definitions for Sepsis and Septic Shock; ICU, intensive care unit.

Predictive performance comparison across temporal sampling densities

In the MIMIC-IV development set with 5-fold cross-validation, the 2-point sampling group (0 h and 23 h) achieved optimal discrimination performance, with an AUROC of 0.814 (95% CI 0.809–0.818) and an AUPRC of 0.537 (95% CI 0.528–0.546). Compared with the single-point baseline group, the 2-point baseline group showed improvements (0.122, from 0.692 to 0.814; $P<0.001$). Compared with the 24-point hourly discrete sampling group, the AUROC difference in the 2-point group was not statistically significant (0.814 vs. 0.814, $P=0.961$), and the AUPRC was slightly lower (0.537 vs. 0.542) without statistical significance, indicating that within a fixed 24-hour observation window, the simplified two-point sampling strategy achieved predictive equivalence to hourly dense monitoring.

In terms of the performance gradient, the AUROC increased sharply from 0.692 in the 1-point group to 0.814 in the 2-point group and then plateaued. The subsequent addition of intermediate sampling timepoints yielded no further improvement, with the AUROC fluctuating in the range of 0.806–0.810 for the 3-point (0.809, $P=0.082$), 4-point (0.806, $P=0.028$), and 6-point (0.810, $P=0.141$) configurations, without significantly surpassing the 2-point group level. This "performance saturation" phenomenon suggests that, after baseline and endpoint information was obtained, intermediate timepoints filled via LOCF did not provide additional genuine predictive signals to the model.

Given the class imbalance characteristic of in-hospital sepsis mortality outcomes, precision–recall analysis provides a more informative evaluation. The 2-point group AUPRC was 0.537, which was close to that of the 24-point group (0.542) and significantly greater than that of the single-point group (0.339) and the intermediate multipoint sampling groups (3-point 0.528, 4-point 0.525, and 6-point 0.528). For the Brier score, the 2-point group value of 0.182 was essentially equivalent to the 24-point group value of 0.186, further supporting the reliability of the two-point strategy in terms of probabilistic prediction accuracy.

The detailed numerical results for all temporal sampling strategies are summarized in Table 3, confirming that the 2-point configuration captures the majority of predictive information with minimal additional benefit from intermediate sampling.

Table 3 Predictive performance of models under different temporal sampling density configurations in the MIMIC-IV development cohort

Configuration	Points	Sampling strategy	AUROC (95% CI)	AUPRC (95% CI)	P value
1-point	1	{0}	0.692*** (0.686–0.698)	0.339 (0.331–0.347)	<.001
2-point	2	{0, 23}	0.814 (0.809–0.818)	0.537 (0.528–0.546)	Ref.
3-point	3	{0, 12, 23}	0.809 (0.806–0.812)	0.528 (0.519–0.537)	0.082
4-point	4	{0, 8, 16, 23}	0.806* (0.802–0.810)	0.525 (0.517–0.533)	0.028
6-point	6	{0, 5, 10, 14, 19, 23}	0.810 (0.805–0.815)	0.528 (0.516–0.540)	0.141
24-point	24	Hourly (0–23)	0.814 (0.808–0.820)	0.542 (0.533–0.551)	0.961

AUROC values are presented with 95% confidence intervals derived from 1,000 bootstrap resamples. Asterisks indicate Bonferroni-corrected P values from pairwise DeLong's test compared with the 2-point [0–23 h] strategy (*P<0.05, ***P<0.001).

Abbreviations: AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; CI, confidence interval; LSTM, long short-term memory; Ref., reference group.

405

406 Figure 3A illustrates this performance saturation phenomenon: the AUROC increased sharply
407 from 0.692 in the 1-point group to 0.814 in the 2-point group, representing an increase of 0.122—
408 the largest performance gain observed across all adjacent group comparisons; thereafter, the
409 AUROC for the 3-point through 6-point configurations fluctuated within the narrow 0.806–0.810
410 range, and the 24-point group plateaued at 0.814, indicating that predictive performance entered a
411 saturation plateau beginning at the 2-point configuration. Figure 3B presents the forest plot of
412 AUROC differences relative to the 2-point reference group with corresponding 95% confidence
413 intervals, corroborating that, apart from the single-point group ($\Delta\text{AUROC} = -0.122$, $P < 0.001$),
414 the magnitude of differences for all remaining groups fell below 0.01, with the 24-point group
415 essentially overlapping the null reference line ($P = 0.961$). Collectively, these findings demonstrate
416 that the optimal two-point sampling strategy sufficiently captures the critical clinical information
417 governing 24-hour disease evolution in elderly sepsis patients, achieving predictive equivalence
418 to the 24-point hourly discrete sampling strategy without additional predictive value conferred by
419 intermediate timepoint acquisition.

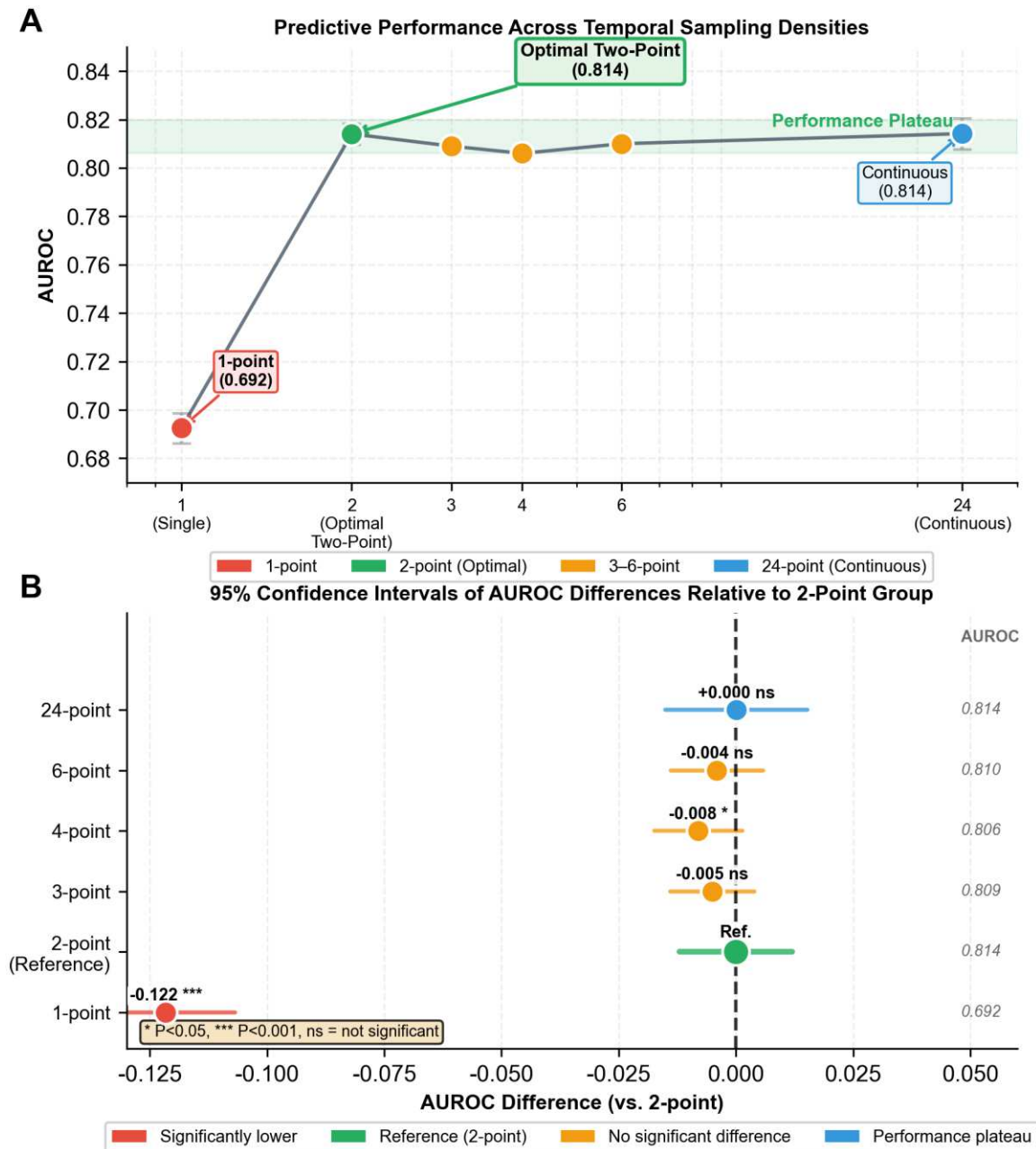


Figure 3 Performance saturation analysis of temporal sampling strategies. (A) Test set AUROC values (with 95% confidence intervals) for the LSTM models trained with 1- to 6-point sampling strategies. The dashed vertical line indicates the inflection point where adding additional sampling timepoints no longer yields a statistically significant improvement in discriminative performance. (B) Forest plot showing pairwise comparisons relative to the 2-point [0–23 h] reference strategy via DeLong's test. The error bars represent 95% confidence intervals for

AUROC differences. All P values were Bonferroni-corrected with a significance threshold of $\alpha=0.0033$ (indicated by the dashed vertical line). AUROC, area under the receiver operating characteristic curve; LSTM, long short-term memory.

External validation of the optimal sampling strategy

The eICU-CRD database included 11039 samples for external validation of the 2-point sampling model. The model demonstrated robust predictive performance, with an AUROC of 0.763 (95% confidence interval [CI] 0.757–0.769), a Brier score of 0.217, and an AUPRC of 0.546 (95% CI 0.536–0.556). Other performance metrics included an accuracy of 0.714, a recall of 0.382, a precision of 0.802, an F1 score of 0.518, and a specificity of 0.910. The external validation performance of all temporal sampling density configurations is detailed in Table 4. Notably, the AUROC difference between the 2-point and 24-point groups in the external validation was not statistically significant (0.763 vs. 0.768, $P=0.31$), further supporting the equivalence between the two-point strategy and hourly dense monitoring across different healthcare systems.

Table 4 Performance comparison of temporal sampling density configurations in the eICU-CRD external validation cohort

Configuration	AUROC (95% CI)	AUPRC (95% CI)	P value
1-point	0.665 (0.658–0.672)	0.358 (0.348–0.368)	<0.001
2-point	0.763 (0.757–0.769)	0.546 (0.536–0.556)	Ref.
3-point	0.761 (0.755–0.767)	0.542 (0.532–0.552)	0.48
4-point	0.760 (0.754–0.766)	0.540 (0.530–0.550)	0.21
6-point	0.763 (0.756–0.770)	0.547 (0.537–0.557)	0.95
24-point	0.768 (0.762–0.774)	0.556 (0.546–0.566)	0.31

The models were trained on the MIMIC-IV development cohort and directly evaluated on the eICU-CRD cohort without retraining. P values were derived from pairwise DeLong's test against the 2-point [0–23 h] strategy (reference).

Abbreviations: AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; CI, confidence interval; eICU-CRD, eICU Collaborative Research Database; Ref., reference group.

Calibration and clinical utility assessment of the optimal temporal sampling scheme

Beyond discrimination, the accuracy of model output probabilities is another key determinant of clinical deployability. Figure 4 presents the calibration performance of the 2-point model in the MIMIC-IV internal development cohort and the eICU-CRD external validation cohort. The internal cohort metrics were calculated on the basis of out-of-fold predictions from 5-fold cross-validation. The model demonstrated acceptable calibration in both cohorts, with calibration curves distributed approximately along the ideal diagonal (Figures 4A and 4B). The Brier score was 0.182 in the internal validation cohort and 0.217 in the external validation cohort, indicating good probabilistic prediction accuracy in the internal cohort and a modest decrease in calibration in the external cohort without severe systematic risk overestimation or underestimation.

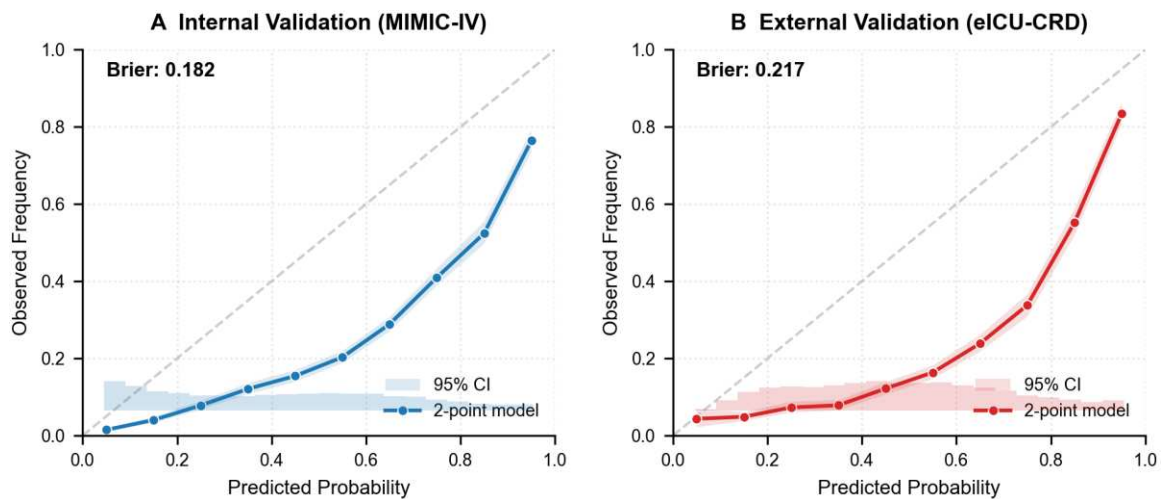


Figure 4 Calibration curves of the 2-point model in the internal validation cohort (MIMIC-IV) (A) and external validation cohort (eICU-CRD) (B). The diagonal dashed line represents perfect

calibration. The solid green line with a shaded area represents the model's predicted probability versus the observed frequency with a 95% confidence interval. The model demonstrated acceptable calibration in both cohorts, with Brier scores of 0.182 and 0.217 in the internal and external validation cohorts, respectively. Abbreviations: MIMIC-IV, Medical Information Mart for Intensive Care IV; eICU-CRD, Electronic ICU Collaborative Research Database; 2-point model, baseline at 0 h and 23 h; Brier, Brier score; CI, confidence interval.

Finally, decision curve analysis (DCA) [38] was used to evaluate the clinical utility of different models in real-world settings. Within the core clinical decision threshold range for ICU sepsis death risk intervention (0.05–0.30), the 2-point model demonstrated a consistent and significant clinical net benefit over both the 1-point baseline model and the "treat all/treat none" benchmark strategies in both the internal and external cohorts. Most notably, throughout the entire decision interval, the net benefit curve of the 2-point group essentially overlapped with that of the 24-point hourly discrete sampling group (Figures 5A and 5B). This result supports the core clinical hypothesis of this study: the simplified two-point data extraction strategy can provide clinical decision support comparable to that of 24-point hourly discrete sampling while substantially reducing the data collection burden on clinical staff.

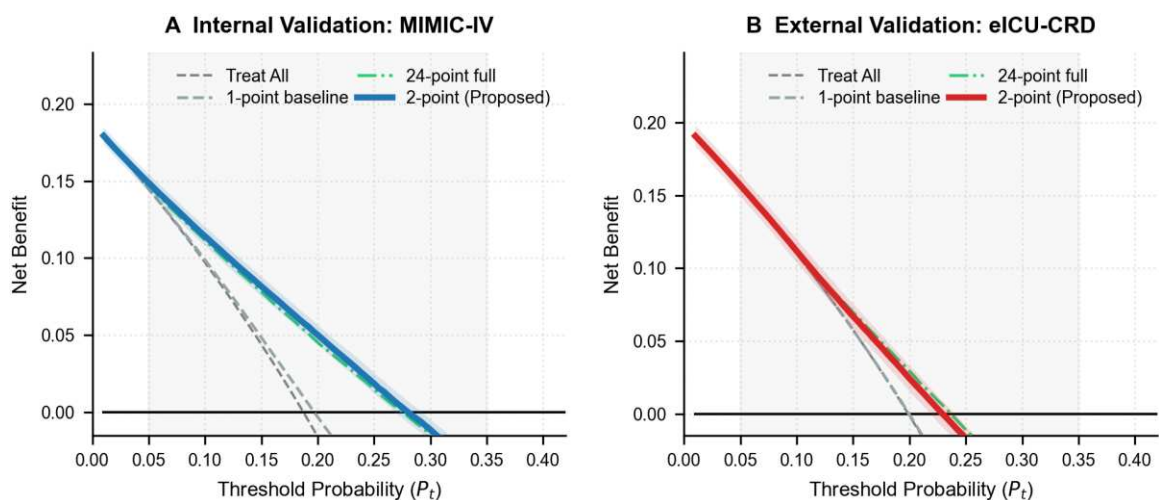


Figure 5 Decision curve analysis (DCA) of the 2-point model for in-hospital mortality prediction in elderly sepsis patients. (A) MIMIC-IV development cohort. (B) The eICU-CRD external validation cohort. Within the core clinical decision threshold range for ICU sepsis death risk intervention (0.05–0.30), the 2-point model demonstrated consistent and significantly greater net clinical benefit than both the 1-point baseline model and the "treat all/treat none" benchmark strategies in both cohorts. The net benefit curve of the 2-point group essentially overlapped with that of the 24-point hourly discrete sampling group throughout the entire decision interval. Abbreviations: DCA, decision curve analysis; MIMIC-IV, Medical Information Mart for Intensive Care IV; eICU-CRD, Electronic ICU Collaborative Research Database; 2-point model, baseline at 0 h and 23 h; 1-point baseline, static baseline at 0 h; 24-point full, hourly discrete sampling from 0--23 h.

Interpretability of the optimal model

To reveal the decision mechanism of the optimal two-point sampling model, SHapley Additive exPlanations (SHAP) [39] analysis was performed on prediction results from representative folds of the MIMIC-IV development set. Global feature importance ranking revealed that 23-hour follow-up features dominated the predictive contribution, whereas only a few 0-hour baseline features ranked among the top contributors, and among static features, only age had a substantial

impact, as illustrated in Figure 6.

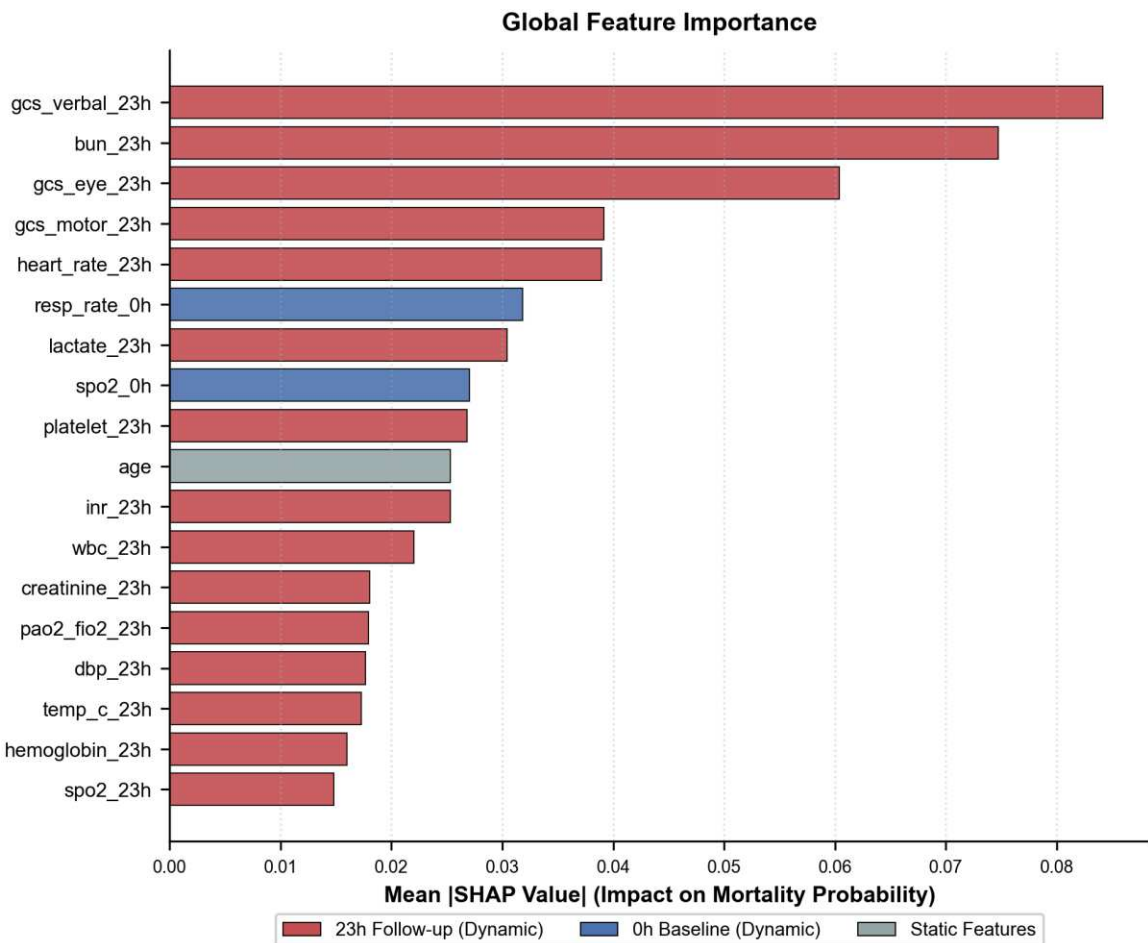


Figure 6 SHAP global feature importance ranking for 2-point model in-hospital mortality prediction in elderly ICU sepsis patients. The horizontal bar chart displays the mean absolute SHAP value for each feature, ranked from highest to lowest impact on mortality probability. The feature colors indicate measurement timepoints: the red bars represent 23-hour follow-up dynamic features, the blue bars represent 0-hour admission baseline dynamic features, and the gray bars represent static demographic features. The GCS verbal response score at 23 hours (gcs_verbal_23h) presented the highest mean |SHAP| value, followed by the blood urea nitrogen score at 23 hours (bun_23h) and the GCS eye opening score at 23 hours (gcs_eye_23h). Among the baseline features, only the admission respiratory rate (resp_rate_0 h) and peripheral oxygen saturation (spo2_0 h) ranked within the top ten. SHAP, SHapley Additive exPlanations; GCS,

Glasgow Coma Scale; bun, blood urea nitrogen; gcs_eye, Glasgow Coma Scale eye opening; gcs_motor, Glasgow Coma Scale motor response; heart_rate, beats per minute; resp_rate, respiratory rate; spo2, peripheral oxygen saturation; inr, international normalized ratio; wbc, white blood cell count; pao2_fio2, ratio of partial pressure of arterial oxygen to fraction of inspired oxygen; dbp, diastolic blood pressure; temp_c, temperature in Celsius; spo2_23h, peripheral oxygen saturation at 23 hours.

Specifically, the 23-hour GCS verbal response score (gcs_verbal_23h) ranked first with the highest mean absolute SHAP value, followed by 23-hour BUN (bun_23h), 23-hour GCS eye opening score (gcs_eye_23h), 23-hour heart rate (heart_rate_23h), and 23-hour lactate (lactate_23h). The 23-hour GCS motor response score (gcs_motor_23h) and 23-hour INR (inr_23h) also ranked within the top seven. Among the 0-hour baseline features, only the admission respiratory rate (resp_rate_0 h) and SpO₂ (spo2_0 h) entered the top ten. Among the static features, only age ranked 10th, with the remaining static features contributing little. This distribution pattern indicates that the 24-hour treatment response status after ICU admission—particularly in terms of neurological function and metabolic indicators—has greater prognostic value than the baseline status at admission.

Single-feature dependence analysis further revealed the clinical behavior patterns of key features. The SHAP dependence plot of the 23-hour GCS score showed clear stepwise risk stratification (Figure 7): patients with a score of 1 (no response) had a median SHAP value of approximately 0.11, indicating an extremely high death risk contribution; at a score of 2 (incomprehensible sounds), the risk contribution decreased to approximately 0.085; at a score of 3 (inappropriate words), it further decreased to approximately 0.035; and at scores of 4 (confused) and 5 (oriented), the SHAP values became negative (approximately -0.03 and -0.075, respectively), indicating that patients with intact consciousness had significantly reduced death risk. Each one-grade improvement in the GCS score was associated with an approximately linear one-step

decrease in the risk of death, confirming the central role of neurological function assessment in determining the prognosis of elderly sepsis patients.

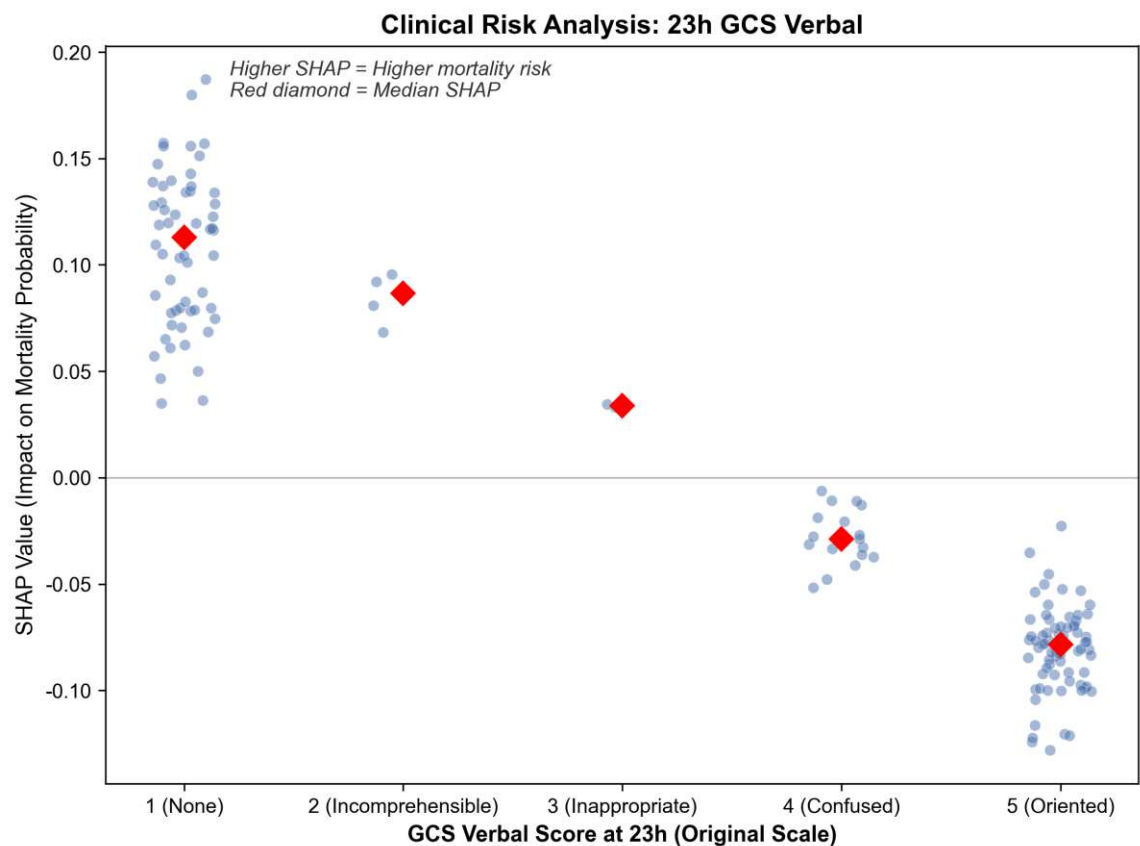


Figure 7 SHAP dependence plot of the 23-hour Glasgow Coma Scale (GCS) verbal score for in-hospital mortality prediction in elderly sepsis patients. The x-axis represents the original GCS verbal score (1 = no response, 2 = incomprehensible sounds, 3 = inappropriate words, 4 = confused, 5 = oriented), and the y-axis represents the SHAP value (impact on mortality probability). Red diamonds denote the median SHAP value at each score level. The plot reveals a clear stepwise risk stratification pattern: each one-grade improvement in the GCS score is associated with an approximately linear one-step decrease in the risk of death, confirming the central role of neurological function assessment in determining prognosis. Abbreviations: SHAP, SHapley Additive ExPlanations; GCS, Glasgow Coma Scale.

The SHAP dependence plot for 23-hour lactate, illustrated in Figure 8, revealed a nonlinear threshold effect. When the lactate concentration was less than 2.0 mmol/L, the SHAP values were generally negative, suggesting a protective effect of low lactate levels; when the lactate concentration exceeded 2.0 mmol/L, the SHAP values shifted from negative to positive, and the death risk contribution began to increase monotonically; and when the lactate concentration exceeded 4.0 mmol/L, the risk contribution rose sharply. This finding is highly consistent with the definition of 2.0 mmol/L lactate as the critical threshold for tissue perfusion inadequacy in sepsis clinical guidelines, providing physiological evidence supporting the clinical interpretability of model predictions.

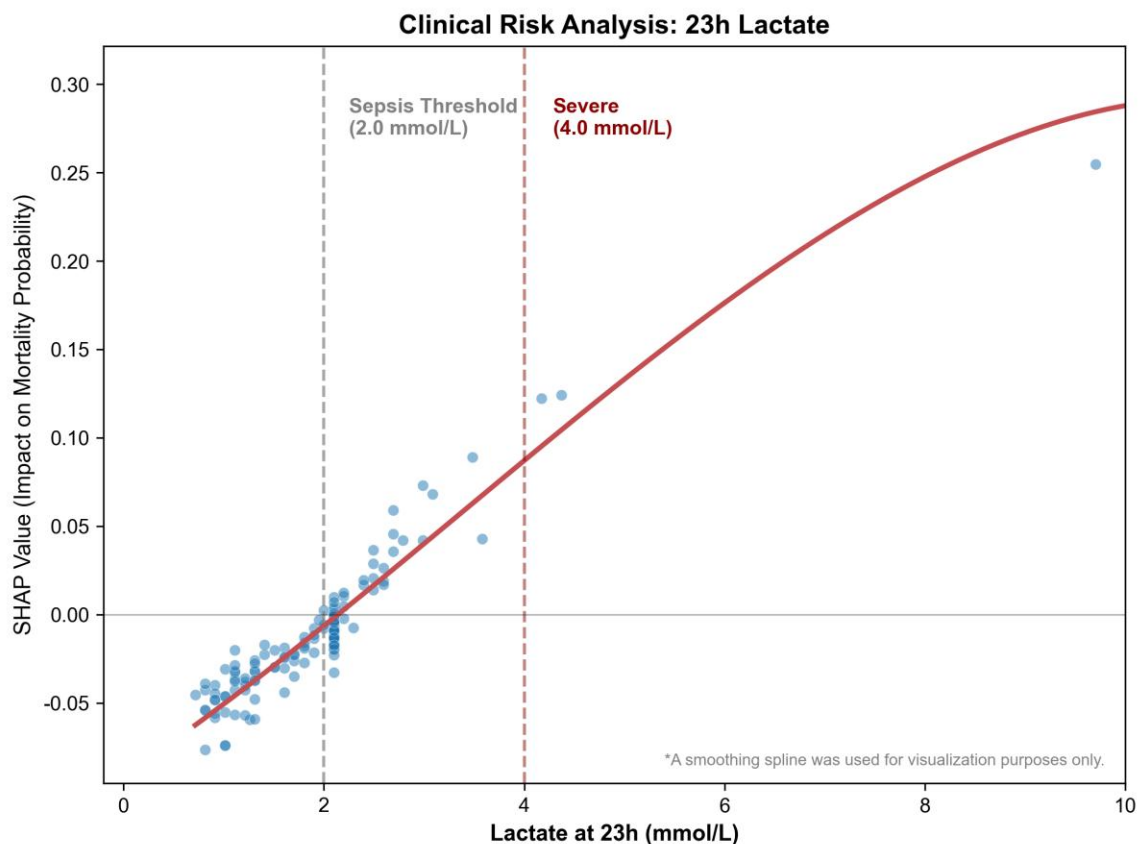


Figure 8 SHAP dependence plot of 23-hour lactate for in-hospital mortality prediction in elderly sepsis patients. The x-axis represents the lactate value at 23 hours (mmol/L), and the y-axis represents the SHAP value (impact on mortality probability). A smoothing spline was fitted for

visualization purposes. When lactate is below 2.0 mmol/L (the clinical threshold for tissue perfusion inadequacy in sepsis guidelines), SHAP values are generally negative, suggesting a protective effect; when lactate exceeds 2.0 mmol/L, SHAP values shift from negative to positive, and the death risk contribution begins to increase monotonically; and when lactate exceeds 4.0 mmol/L, the risk contribution increases sharply. Abbreviations: SHAP, SHapley Additive exPlanations.

Sensitivity analysis

To verify that the performance advantage of the optimal two-point sampling scheme (0 h and 23 h) did not depend solely on the chance selection of specific timepoints, this study fixed the 0-hour baseline timepoint and systematically used 6-hour, 12-hour, 18-hour, and 23-hour timepoints as the second timepoint to construct control groups, evaluating the effect of second-sampling-point selection on model predictive performance.

The results revealed that as the second timepoint shifted from 6 hours to 23 hours, the model AUROC increased monotonically: the 0-hour and 6-hour combination yielded an AUROC of 0.736 (95% CI 0.728–0.743) and an AUPRC of 0.398; the 0-hour and 12-hour combination increased to an AUROC of 0.774 (95% CI 0.769–0.779) and an AUPRC of 0.460; the 0-hour and 18-hour combination further increased to an AUROC of 0.796 (95% CI 0.789–0.803) and an AUPRC of 0.499; and the 0-hour and 23-hour combination (the optimal scheme) achieved the highest AUROC of 0.814 (95% CI 0.809–0.816) and an AUPRC of 0.537. The Brier score improved synchronously, decreasing from 0.205 for the 6-hour combination to 0.175 for the 23-hour combination. These detailed performance metrics for all second-timepoint configurations are presented in Table 5, highlighting the monotonic increase in predictive accuracy as the second timepoint approaches the 24-hour endpoint and supporting the clinical rationale for selecting 24 hours after ICU admission as the key assessment timepoint.

Table 5 Effect of second sampling timepoint selection on model performance (sensitivity analysis)

Configuration	AUROC (95% CI)	AUPRC	F1-score	Brier
1-point (0 h)	0.692 (0.684–0.698)	0.339	0.372	0.232
2-point (0, 6 h)	0.736 (0.728–0.743)	0.398	0.434	0.205
2-point (0, 12 h)	0.774 (0.769–0.779)	0.460	0.469	0.191
2-point (0, 18 h)	0.796 (0.789–0.803)	0.499	0.502	0.179
2-point (0, 23 h)	0.814 (0.809–0.816)	0.537	0.512	0.175

All the configurations used an identical fixed-architecture LSTM; only the second sampling timepoint varied.

Abbreviations: AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; LSTM, long short-term memory.

Discussion

This study systematically compared the effects of six temporal sampling density strategies on in-hospital mortality prediction performance in elderly sepsis patients via a fixed-architecture LSTM network. The core findings were as follows: from the admission static baseline (1-point group AUROC 0.692) to the simplified two-point sampling (2-point group AUROC 0.814), the predictive performance improvement reached 0.122—a significant increase from the two-point group to the 24-point group; the simplified two-point sampling strategy (admission baseline and 24-hour treatment endpoint) achieved predictive equivalence to hourly dense monitoring (24-point group AUROC 0.814, $P=0.961$). After baseline and endpoint information, further addition of intermediate timepoints (3-point, 4-point, and 6-point) yielded no additional predictive gain, with the AUROC fluctuating within the 0.806–0.810 range, suggesting a performance saturation phenomenon within the 24-hour observation window. External validation in the eICU-CRD database further confirmed the equivalence between the two-point strategy and hourly dense monitoring (AUROC

0.763 vs. 0.768, $P=0.31$), corroborating the clinical applicability of this simplified sampling approach across different healthcare systems.

This "performance saturation" phenomenon can be interpreted in relation to the literature at three levels. At the level of sampling frequency exploration, Rasmussen et al. [8] compared three discrete granularities (1 hour, 4, and 24) in which higher-frequency data improved performance but increased the computational burden; however, their granularity gradient was too sparse to reveal the pattern of diminishing marginal returns for intermediate sampling. Zargoush et al. [24] systematically evaluated the timeliness of historical information and reported that stale data at intermediate timepoints might impair performance—a direction consistent with our finding of "no additional gain from adding intermediate points"—but neither study conducted a systematic exploration from the perspective of a sampling frequency gradient. This study is the first to compare the complete sampling density gradient from 24°C within a fixed 24-hour window and to reveal the critical density at which performance saturates. At the level of clinical deployment of simplified sampling, Desautels et al. [20] demonstrated with the InSight system that effective prediction can be achieved using only minimal variables such as vital signs, SpO₂, GCS, and age (AUROC 0.781). This study advances the simplified approach from "which features" to "which timepoints" and, through SHAP interpretability analysis, reveals the core driving features at key timepoints (e.g., 23-hour GCS verbal score, BUN, and lactate), thereby advancing the simplified approach from the independent dimensions of "which features are needed" and "at which timepoints data should be collected" to the precise positioning of "which indicators should be prioritized at key assessment nodes."

To contextualize the discriminative performance of our model within existing clinical benchmarks, we compared it with the performance of traditional ICU severity scores in elderly sepsis populations reported in the literature. Independent validation studies in

elderly or very elderly sepsis patients revealed AUROC values for the APACHE II score ranging from 0.6890.784, for the SOFA score from 0.6880.842, and for the APACHE IV score of approximately 0.722 in patients over 80 years of age [40–42]. The simplified two-point LSTM model proposed in this study achieved an AUROC of 0.814 in the MIMIC-IV development set, which is at the upper limit of the range reported for traditional scores; even in the external validation set (eICU-CRD), where it decreased to 0.763, its performance remained comparable to that of most traditional scores in similar populations. Importantly, traditional scoring systems typically rely on continuous collection and complex weighting of multiple physiological indicators within 24 hours of intensive care unit (ICU) admission, whereas this model is equivalent to or even superior in discriminative performance, with only sparse data from two timepoints (admission baseline and 24-hour endpoint), highlighting its practical advantage for clinical deployment.

From a clinical translation perspective, the two-point sampling strategy offers substantial practical advantages. Data collection at only two timepoints greatly reduces the clinical workload, aligning with the core demand for "simplified usability" of predictive tools in high-frequency, high-pressure ICUs [21,39]. Second, the two-point model demonstrated high specificity (0.910) in the external validation, which has direct clinical value for alleviating the alarm fatigue problem prevalent in ICUs [43]. Liu et al. [43] noted in a commentary published in Lancet Digital Health that high sensitivity with concomitant low specificity is the core mechanism underlying clinical alarm fatigue, calling for substantial improvement in specificity while maintaining basic sensitivity to gain clinicians' trust. The specificity of our two-point model (0.910) directly addresses this clinical pain point. Finally, SHAP interpretability analysis revealed that dynamic changes in key features such as the 23-hour GCS verbal score, BUN, and lactate are the core drivers of prediction, which not only enhances the clinical credibility of the model [17,18] but also provides data-driven

validation for the traditional clinical experience that 24-hour treatment response assessment serves as a critical prognostic node.

Regarding the mechanism by which intermediate multipoint sampling fails to yield additional gain, we hypothesize that the primary reason lies in the noise amplification effect of the LOCF forward-fill strategy in a data-sparse environment. This study calculated the raw missingness rates of dynamic features at each timepoint (Supplementary Material 2). The results revealed two fundamentally different acquisition modes for ICU temporal data: "streaming" vital signs (heart rate, blood pressure, oxygen saturation, etc.) continuously recorded by monitors, with extremely low missing rates ($<6\%$), and "pulsed" laboratory indices (lactate, creatinine, INR, etc.) typically collected only 1–2 times daily, with instantaneous missing rates at any single-hour slice reaching 84–95%, and GCS missing rates also reaching 61–75%. In the two-point scheme, only the 0-hour and 23-hour timepoints require extraction of genuine observations, minimizing LOCF fill instances; in the 3-point, 4-point, and 6-point schemes, the data demand at intermediate timepoints increased substantially, but the actual collection frequency of laboratory indices was far below the preset sampling density, forcing LOCF to repeatedly copy stale observations and inject artificial artifacts into the model. Moor et al. [26] explicitly noted in their MGP-AttTCN model that traditional LOCF imputation introduces noise and dilutes genuine physiological change signals in ICU temporal data. This study provides new empirical evidence for this mechanism from the perspective of sampling frequency, suggesting that, under the realistic constraint of ICU data sparsity, a "less but precise" timepoint selection strategy is superior to a "more but noisy" dense sampling scheme. We chose LOCF over more advanced imputation methods such as MICE or KNN, primarily because it directly mirrors the clinical workflow (clinicians typically carry forward the most recent known value until new data are generated) and offers the advantages of zero hyperparameters and extreme computational simplicity, making it more suitable for bedside deployment in real-time ICU warning systems. Nevertheless, we acknowledge that the phenomenon of intermediate sampling points yielding no additional gain

may partially reflect artificial artifacts introduced by LOCF, and future research should compare alternative imputation strategies within the same framework to verify the robustness of this conclusion.

The methodological design of this study has several distinctive features. The use of a fixed LSTM architecture, a unified feature set (23 predictors selected on the basis of established sepsis prognostic literature, clinical routine accessibility, and collinearity exclusion principles; see Supplementary Material 1 for details), and a consistent training strategy ensured that the only difference between the six experimental groups was temporal sampling density, thereby precisely isolating the independent effect of this factor on predictive performance. Focusing on the high-risk subgroup of elderly sepsis patients aged ≥ 65 years addresses both the practical needs of ICU clinical practice in the context of global population aging and the greater physiological trajectory heterogeneity in elderly patients, enabling the effect of temporal sampling density to be fully manifested [35,44]. Seymour et al. [45] identified four clinical phenotypes of sepsis (α , β , γ , δ) from large ICU cohorts via machine learning clustering analysis, in which the β phenotype was characterized by advanced age and chronic disease burden and the δ phenotype exhibited the highest mortality, revealing substantial heterogeneity in the risk of death among sepsis patients. This heterogeneity is precisely the key foundation enabling temporal models to surpass static prediction.

However, this study has several limitations. First, the retrospective design cannot establish causal relationships; Sepsis-3 diagnosis relies on retrospective ICD coding and SOFA score changes, which may introduce identification bias; and model performance requires further validation in independent prospective cohorts [47]. Second, both databases are from the U.S. healthcare system, and the generalizability of the model to other countries or regions remains to be confirmed [19]. Third, this study did not incorporate treatment intervention information (e.g., antibiotic use, vasoactive medications), and future models could be further optimized by integrating treatment dynamics. Finally, the optimality of the two-point scheme was based on the

24-hour in-hospital death prediction task; optimal sampling strategies may differ for other clinical scenarios (e.g., early sepsis recognition or longer observation windows).

Conclusions

This study systematically compared the effects of six temporal sampling density strategies on in-hospital mortality prediction performance in elderly sepsis patients via a fixed-architecture LSTM network. We found that the simplified two-point sampling strategy (admission baseline and 24-hour treatment endpoint) achieved predictive equivalence to hourly dense monitoring (AUROC 0.814 vs. 0.814, $P=0.961$), whereas adding intermediate sampling timepoints did not provide additional predictions. External validation in the eICU-CRD database further confirmed the equivalence between the two-point strategy and hourly dense monitoring (AUROC 0.763 vs. 0.768, $P=0.31$), corroborating the cross-robustness of this simplified sampling approach. SHAP interpretability analysis revealed that 24-hour treatment response status—particularly the GCS verbal score, BUN, and lactate—has greater prognostic value than does admission baseline status. This study provides an efficient data acquisition approach for the clinical deployment of ICU sepsis mortality prediction models, maintaining high predictive accuracy and clinical credibility while reducing the clinical workload. Prospective studies are warranted to further validate these findings and explore the differential selection of optimal temporal sampling strategies across different clinical scenarios.

Abbreviations

AUROC: Area under the receiver operating characteristic curve; AUPRC: Area under the precision–recall curve; BUN: Blood urea nitrogen; CI: Confidence interval; DCA: Decision curve analysis; eICU-CRD: Electronic ICU Collaborative Research Database; GCS: Glasgow Coma Scale; ICU: Intensive Care Unit; INR: International normalized ratio; LOCF: Last observation carried forward; LSTM: Long short-term memory; ML: Machine learning; MIMIC-IV: Medical

Information Mart for Intensive Care IV; Sepsis-3: Third International Consensus Definitions for Sepsis and Septic Shock; SHAP: SHapley Additive exPlanations; SOFA: Sequential Organ Failure Assessment.

Declarations

Ethics approval and consent to participate

Because the MIMIC-IV and eICU-CRD databases used in this study have been strictly deidentified in accordance with the Safe Harbor provisions of the Health Insurance Portability and Accountability Act (HIPAA), no additional institutional review board (IRB) authorization was needed. The research team completed mandatory data use training before obtaining data access privileges. In accordance with the established policies of the database administrators, informed consent was waived for this type of research.

Consent for publication

Not applicable.

Availability of data and materials

The MIMIC-IV (version 3.1) and eICU-CRD (version 2.0) datasets analyzed in the present study are publicly available through the MIMIC-IV database (<https://physionet.org/content/mimiciv/3.1/>) and eICU-CRD database (<https://physionet.org/content/eicu-crd/2.0/>) after completion of the required credentialing training and data use agreement. There are no restrictions to their use by nonacademics.

ts

The authors declare that they have no competing interests.

Funding

754 Not applicable.

755 **Authors' contributions**

756 WG contributed to study design, data analysis, and model development. JY and YX performed
757 data eAP interpretability analysis. BL supervised the study and revised the manuscript. All
758 authors read and approved the final manuscript.

759 **Acknowledgments**

760 Not applicable.

761 **Code availability**

762 To ensure reproducibility, the complete source code developed for this study, along with the
763 trained best-performing model and associated submodels, has been made publicly available. The
764 repository is hosted at <https://github.com/frankfrank6/sepsis-temporal-sampling>. Researchers can
765 reproduce the experimental results via the provided loading commands with a Python runtime
766 environment configured with PyTorch (version 1.12.0).

767 **Supplementary Information**

768 Supplementary Material 1. Candidate variable pool and feature selection rationale.

769 Supplementary Material 2. Missing rates of dynamic features by time point.

770

771 **References**

772 1. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions
773 for sepsis and septic shock (sepsis-3). *JAMA*. 2016;315(8):801-10.
774 doi:10.1001/jama.2016.0288.

- 775 2. Rudd KE, Johnson SC, Agesa KM, et al. Global, regional, and national sepsis incidence and
776 mortality, 1990-2017: analysis for the global burden of disease study. *Lancet*.
777 2020;395(10219):200-11. doi:10.1016/S0140-6736(19)32989-7.
- 778 3. Torvik MA, Nymo SH, Bjørnsen LP, et al. Patient characteristics in sepsis-related deaths:
779 prevalence of advanced frailty, comorbidity, and age in a Norwegian hospital trust.
780 *Infection*. 2023;51(4):1103-15. doi:10.1007/s15010-023-02035-z.
- 781 4. Zhu XY, Jiang ZM, Li X, et al. Interpretive machine learning predicts short-term mortality risk
782 in elderly sepsis patients. *Front Physiol*. 2025;16:1549138.
783 doi:10.3389/fphys.2025.1549138.
- 784 5. Dai M, et al. Association of lactate-to-albumin ratio with 28-day mortality in elderly sepsis
785 patients: a retrospective MIMIC-IV database analysis. *BMC Infect Dis*. 2026;.
786 doi:10.1186/s12879-026-12665-4.
- 787 6. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease
788 classification system. *Crit Care Med*. 1985;13(10):818-29.
- 789 7. Vincent JL, Moreno R, Takala J, et al. The SOFA (sepsis-related organ failure assessment)
790 score to describe organ dysfunction/failure. *Intensive Care Med*. 1996;22(7):707-10.
- 791 8. Rasmussen LS, Lundbye-Christensen S, Thiesson B, et al. Dynamic and explainable machine
792 learning prediction of mortality in patients in the intensive care unit: a retrospective study of
793 high-frequency data in electronic patient records. *Lancet Digit Health*. 2023;5(7):e499-
794 e508. doi:10.1016/S2589-7500(23)00112-3.
- 795 9. Li H. Long-term outcomes in older patients with sepsis in the ICU: a retrospective study.
796 *Medicine (Baltimore)*. 2022;101(42):e31251. doi:10.1097/MD.00000000000031251.

- 797 10. Zhang G, Shao F, Yuan W, et al. Predicting sepsis in-hospital mortality with machine
798 learning: a multicenter study using clinical and inflammatory biomarkers. *Eur J Med Res*.
799 2024;29:156.
- 800 11. Deng Y, Liu S, Wang Z, et al. Explainable time-series deep learning models for the prediction
801 of mortality, prolonged length of stay and 30-day readmission in intensive care patients.
802 *Front Med*. 2022;9:933037. doi:10.3389/fmed.2022.933037.
- 803 12. Shen L, Wu J, Lan J, et al. Interpretable machine learning-based prediction of 28-day
804 mortality in ICU patients with sepsis: a multicenter retrospective study. *Front Cell Infect*
805 *Microbiol*. 2025;14:1500326. doi:10.3389/fcimb.2024.1500326.
- 806 13. Gao J, Lu Y, Ashrafi N, et al. Prediction of sepsis mortality in ICU patients using machine
807 learning methods. *BMC Med Inform Decis Mak*. 2024;24:228.
- 808 14. Yu Z, Ashrafi N, Li H, et al. Prediction of 30-day mortality for ICU patients with sepsis-3.
809 *BMC Med Inform Decis Mak*. 2024;24(1):223.
- 810 15. Cheng YW, Kuo PC, Chen SH, et al. Early prediction of mortality at sepsis diagnosis time in
811 critically ill patients by using interpretable machine learning. *J Clin Monit Comput*.
812 2024;38:271-9.
- 813 16. Nikravangolsefid N, Reddy S, Truong HH, et al. Machine learning for predicting mortality in
814 adult critically ill patients with sepsis: a systematic review. *J Crit Care*. 2024;84:154889.
815 doi:10.1016/j.jcrc.2024.154889.
- 816 17. Oliveira TQ, Carvalho LA, Sousa FRC, et al. Responsible AI for sepsis prediction: bridging
817 the gap between machine learning performance and clinical trust. *J Clin Med*.
818 2026;15(6):2251. doi:10.3390/jcm15062251.

18. Yilmaz Baser H, Evran T, Cifci MA. Machine learning-augmented triage for sepsis: real-time ICU mortality prediction using SHAP-explained meta-ensemble models. *Biomedicines*. 2025;13(6):1449. doi:10.3390/biomedicines13061449.
19. Tranchellini F, Farag Y, Jutzeler C, Meegahapola L. Evaluating deep learning sepsis prediction models in ICUs under distribution shift: a multicenter retrospective cohort study. *medRxiv*. 2025;. doi:10.1101/2025.07.31.25332542.
20. Desautels T, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform*. 2016;4(3):e28.
21. Moor M, Rieck B, Horn M, et al. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Front Med (Lausanne)*. 2021;8:607952. doi:10.3389/fmed.2021.607952.
22. Liu S, Fu B, Wang W, et al. Dynamic sepsis prediction for intensive care unit patients using XGBoost-based model with novel time-dependent features. *IEEE J Biomed Health Inform*. 2022;26(8):4258-69. doi:10.1109/JBHI.2022.3162105.
23. Ul Alam M, Henriksson A, Valik JK, et al. Deep learning from heterogeneous sequences of sparse medical data for early prediction of sepsis. *Comput Methods Programs Biomed*. 2020;105766.
24. Zargoush M, Sameh A, Javadi M, et al. The impact of recency and adequacy of historical information on sepsis predictions using machine learning. *Sci Rep*. 2021;11:20942.
25. Moor M, Horn M, Rieck B, et al. MGP-AttTCN: an interpretable machine learning model for the prediction of sepsis. *Proc Mach Learn Healthc*. 2021;.
26. Schamoni A, Frank T, Fiedler F, et al. Sepsis prediction in intensive care using machine learning and temporal data. *Artif Intell Med*. 2021;118:102104. doi:10.1016/j.artmed.2021.102104.

- 843 27. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for
844 reporting clinical prediction models that use regression or machine learning methods. *BMJ*.
845 2024;385:e078378.
- 846 28. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational
847 studies in epidemiology (STROBE) statement: guidelines for reporting observational
848 studies. *PLoS Med*. 2007;4(10):e296.
- 849 29. Jansen TC, van Bommel J, Schoonderbeek FJ, et al. Early lactate-guided therapy in intensive
850 care unit patients: a multicenter, open-label, randomized controlled trial. *Am J Respir Crit*
851 *Care Med*. 2010;182(6):752-61.
- 852 30. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale.
853 *Lancet*. 1974;304(7872):81-4.
- 854 31. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic
855 comorbidity in longitudinal studies: development and validation. *J Chronic Dis*.
856 1987;40(5):373-83.
- 857 32. Gilbert T, Neuburger J, Kraindler J, et al. Development and validation of a hospital frailty risk
858 score focusing on older people in acute care settings using electronic hospital records: an
859 observational study. *Lancet*. 2018;391(10132):1775-82.
- 860 33. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-80.
861 doi:10.1162/neco.1997.9.8.1735.
- 862 34. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other
863 measures of statistical accuracy. *Stat Sci*. 1986;1(1):54-75.
- 864 35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more
865 correlated receiver operating characteristic curves. *Biometrics*. 1988;44(3):837-45.

- 866 36. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316(7139):1236-8.
867 doi:10.1136/bmj.316.7139.1236.
- 868 37. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing
869 models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat*
870 *Med*. 1996;15(4):361-87.
- 871 38. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction
872 models. *Med Decis Making*. 2006;26(6):565-74. doi:10.1177/0272989X06295361.
- 873 39. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf*
874 *Process Syst*. 2017;30:4765-74.
- 875 40. Tekin B, Yalçın GP, Çevik Y, et al. The comparison of scoring systems: SOFA, APACHE-II,
876 LODS, MODS, and SAPS-II in critically ill elderly sepsis patients. *J Infect Dev Ctries*.
877 2024;18(1):122-30.
- 878 41. Haas LEM, Karakus A, Verstraete R, et al. Performance of the quick SOFA in very old ICU
879 patients admitted with sepsis. *Acta Anesthesiol Scand*. 2020;64(4):508-16.
- 880 42. Nguyen TTM, Tran TTA, Hoang TTV, et al. Sequential organ failure assessment (SOFA)
881 score for predicting mortality in patients with sepsis in Vietnamese intensive care units.
882 *BMJ Open*. 2023;13(3):e064870.
- 883 43. Au-Yeung WM, Sahani AK, Isselbacher EM, Armoundas AA. Reduction of false alarms in
884 the intensive care unit using an optimized machine learning based approach. *NPJ Digit Med*.
885 2019;2:86. doi:10.1038/s41746-019-0160-7.
- 886 44. Mou C, Yang J, Wu Q, et al. Progress in sepsis prediction models: from traditional scoring
887 systems to multimodal intelligence and clinical translation. *Front Med*. 2026;13:1732164.
888 doi:10.3389/fmed.2026.1732164.

889 45. Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment
890 implications of novel clinical phenotypes for sepsis. *JAMA*. 2019;321(20):2003-17.
891 doi:10.1001/jama.2019.5791.

892 46. Jagesar ARJ, Otten M, Dam TA, et al. Comparative performance of intensive care mortality
893 prediction models based on manually curated versus automatically extracted electronic
894 health record data. *Int J Med Inform*. 2024;188:105477.
895 doi:10.1016/j.ijmedinf.2024.105477.

896 47. Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from
897 development to external validation. *BMJ*. 2024;384:e074819. doi:10.1136/bmj-2023-
898 074819.

899

900 **Table 2 Baseline characteristics of elderly sepsis patients in the MIMIC-IV**
901 **development cohort**

Characteristic	Survivors (n=16393)	Nonsurvivors (n=3770)	P value
Demographics			
Age, years	76.18 ± 7.64	77.34 ± 7.84	<0.0001
Diabetes, n (%)	5889 (35.9%)	1335 (35.3%)	0.4877
Heart failure, n (%)	2945 (18.0%)	1025 (27.1%)	<0.0001
Severity scores			
Frailty score	0.29 ± 0.71	0.38 ± 0.79	<0.0001
Is frail, n (%)	1542 (9.4%)	511 (13.5%)	<0.0001
Vital signs			
Heart rate, bpm	84.58 ± 18.67	90.24 ± 20.44	<0.0001

SBP, mmHg	112.00 (100.00–127.00)	109.00 (96.50–125.00)	<0.0001
DBP, mmHg	57.00 (50.00–66.00)	58.00 (49.25–68.00)	0.2317
MBP, mmHg	73.00 (65.00–83.00)	72.00 (63.00–83.00)	<0.0001
Respiratory rate, bpm	18.50 (16.00–22.50)	21.00 (17.48–25.00)	<0.0001
SpO ₂ , %	98.00 (96.00–100.00)	97.00 (95.00–99.83)	<0.0001
Temperature, °C	36.72 (36.39–37.06)	36.67 (36.33–37.11)	0.0109
Neurological (GCS)			
GCS eye opening	2.99 ± 1.19	2.73 ± 1.27	<0.0001
GCS verbal response	3.01 ± 1.85	2.69 ± 1.80	<0.0001
GCS motor response	4.90 ± 1.82	4.53 ± 1.93	<0.0001
Laboratory results			
PaO ₂ /FiO ₂ ratio	271.88 ± 151.43	232.48 ± 148.07	<0.0001
Lactate, mmol/L	1.70 (1.20–2.60)	2.20 (1.40–3.70)	<0.0001
Creatinine, mg/dL	1.10 (0.80–1.70)	1.50 (1.00–2.40)	<0.0001
BUN, mg/dL	24.00 (16.00–39.00)	35.00 (23.00–55.00)	<0.0001
Characteristic	Survivors (n=16393)	Nonsurvivors (n=3770)	P value
Demographics			
Age, years	76.18 ± 7.64	77.34 ± 7.84	<0.0001
Diabetes, n (%)	5889 (35.9%)	1335 (35.3%)	0.4877
Heart failure, n (%)	2945 (18.0%)	1025 (27.1%)	<0.0001
Severity scores			

Frailty score	0.29 ± 0.71	0.38 ± 0.79	<0.0001
Is frail, n (%)	1542 (9.4%)	511 (13.5%)	<0.0001
Vital signs			
Heart rate, bpm	84.58 ± 18.67	90.24 ± 20.44	<0.0001
SBP, mmHg	112.00 (100.00–127.00)	109.00 (96.50–125.00)	<0.0001
DBP, mmHg	57.00 (50.00–66.00)	58.00 (49.25–68.00)	0.2317
MBP, mmHg	73.00 (65.00–83.00)	72.00 (63.00–83.00)	<0.0001
Respiratory rate, bpm	18.50 (16.00–22.50)	21.00 (17.48–25.00)	<0.0001
SpO ₂ , %	98.00 (96.00–100.00)	97.00 (95.00–99.83)	<0.0001
Temperature, °C	36.72 (36.39–37.06)	36.67 (36.33–37.11)	0.0109
Neurological (GCS)			
GCS eye opening	2.99 ± 1.19	2.73 ± 1.27	<0.0001
GCS verbal response	3.01 ± 1.85	2.69 ± 1.80	<0.0001
GCS motor response	4.90 ± 1.82	4.53 ± 1.93	<0.0001
Laboratory results			
PaO ₂ /FiO ₂ ratio	271.88 ± 151.43	232.48 ± 148.07	<0.0001
Lactate, mmol/L	1.70 (1.20–2.60)	2.20 (1.40–3.70)	<0.0001
Creatinine, mg/dL	1.10 (0.80–1.70)	1.50 (1.00–2.40)	<0.0001
BUN, mg/dL	24.00 (16.00–39.00)	35.00 (23.00–55.00)	<0.0001
WBC, ×10 ⁹ /L	11.40 (8.30–15.60)	12.90 (8.80–18.20)	<0.0001
Hemoglobin, g/dL	10.09 ± 1.88	9.89 ± 2.10	<0.0001
Platelet, ×10 ⁹ /L	179.00 (129.00–	183.00 (121.00–	0.9457

	248.00)	264.00)	
INR	1.30 (1.20–1.50)	1.50 (1.20–1.90)	<0.0001

Data are presented as the mean $\pm$ standard deviation, median (interquartile range), or number (%). P values were calculated via the independent t test or Mann–Whitney U test for continuous variables and the chi-square test or Fisher's exact test for categorical variables.

Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; MBP, mean blood pressure; SpO₂, peripheral oxygen saturation; GCS, Glasgow Coma Scale; PaO₂/FiO₂, ratio of partial pressure of arterial oxygen to fraction of inspired oxygen; BUN, blood urea nitrogen; WBC, white blood cell count; INR, international normalized ratio.

Supplementary Files

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

- [SupplementaryMaterial2.docx](#)
- [SupplementaryMaterial1.docx](#)