Enhancing Early Sepsis Prediction with Temporal bias EHRs Data

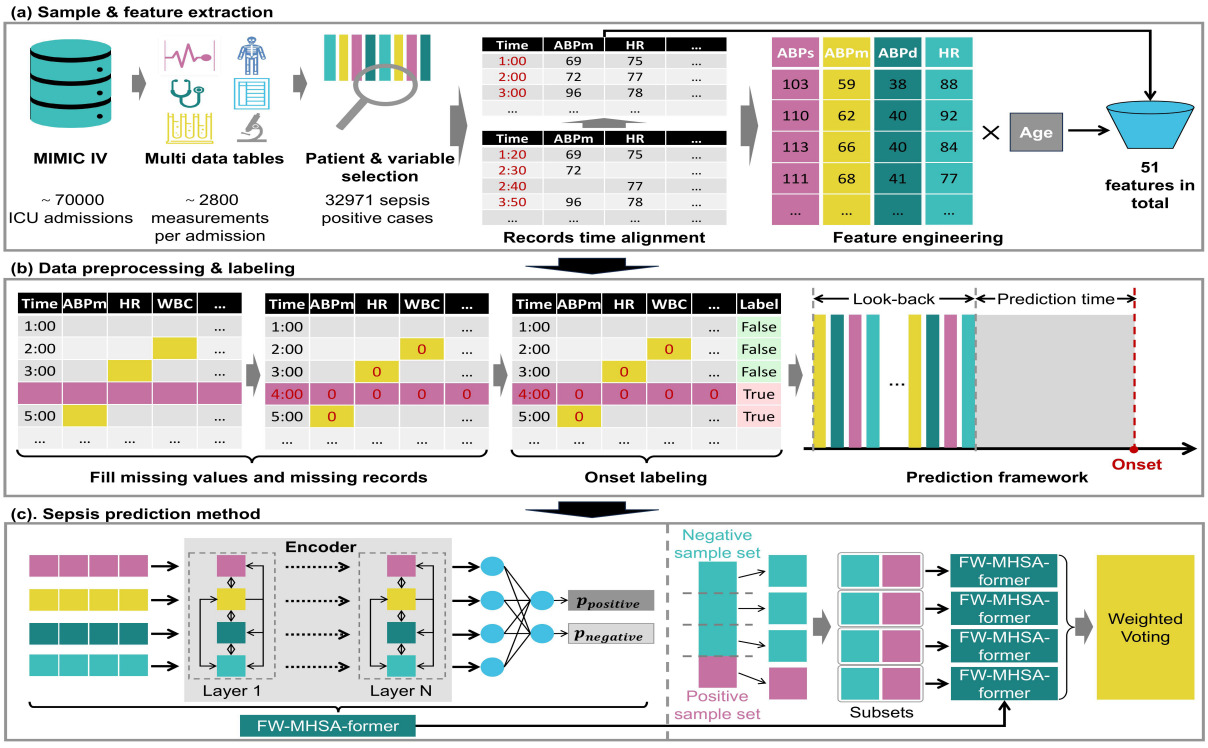
Authors

Mireaye Abudurexiti, Jianshu Wang, Pengfei Zhang, Fangfang Liu, Zhiqiang Jia

Correspondence

12664494@qq.com (Z. Jia)

Graphical Abstract



https://doi.org/10.71321/4t9gsy05

Research Article

Life Conflux

Enhancing Early Sepsis Prediction with Temporal bias EHRs Data

Mireaye Abudurexiti [1†], Jianshu Wang [2†], Pengfei Zhang [1], Fangfang Liu [1], Zhiqiang Jia [1*]

Received: 2026-03-10 | Accepted: 2026-05-18 | Published online: 2026-05-29

Abstract

Background: Sepsis prediction models using electronic health records (EHRs) are often challenged by temporal biases from irregular data entry and severe class imbalance. This study develops a novel deep learning (DL) framework to address these specific challenges for accurate and early sepsis detection.

Methods: We propose a Feature-Wise Multi-Head Self-Attention Transformer (FW-MHSA-former) with an Adaptive Balance-Preserving Ensemble (ABPE). FW-MHSA-former mitigates temporal bias by applying self-attention across medical features to model correlations directly. ABPE resolves class imbalance by partitioning the majority class to train multiple models on balanced datasets, aggregating predictions via weighted voting. The framework was retrospectively validated on the MIMIC-IV dataset; Kaplan-Meier analysis assessed survival outcomes. Visualizing the attention-derived feature correlation matrix enhances interpretability.

Results: The proposed framework achieved a peak Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.94. At this optimal performance, the model demonstrated a recall of 0.90 and an accuracy of 0.87. We compared our method with three classical models and three advanced attention/Transformer-based models. The proposed approach yielded consistently superior performance across all evaluation metrics, including accuracy and F1 score. The KM analysis confirmed that the model effectively stratified patients into high- and low-risk cohorts with statistically significant differences in survival outcomes ($p < 0.001$).

Conclusions: The proposed framework effectively and robustly predicts early sepsis. By addressing timestamp irregularities and class imbalance, it achieves superior accuracy and provides an interpretable tool to enhance clinical decision support in critical care.

Keywords: Sepsis prediction; Transformer; Electronic health records; Deep learning; Interpretable model

Introduction

Sepsis remains a major global health challenge, characterized by high incidence and mortality rates [1-3]. A substantial body of research has established that early and accurate detection is critical for improving patient outcomes [4]. However, the clinical presentation of sepsis often overlaps with other conditions, presenting a diagnostic challenge even for experienced clinicians. Moreover, this challenge is compounded by the underlying immune dysregulation characteristic of sepsis, involving pathways such as cGAS-STING that drive inflammation and organ injury [5]. In response, various screening and predictive scoring systems have been developed, with traditional tools like the Systemic Inflammatory Response Syndrome (SIRS) criteria, Sequential Organ Failure Assessment (SOFA), and quick SOFA (qSOFA) being widely implemented. Despite their utility, these systems exhibit several intrinsic limitations. They primarily rely on static physiological and laboratory parameters, inadequately capturing the complex, non-linear relationships within patient data. Furthermore, their dependence on single-point-in-time measurements fails to account for the dynamic evolution of a patient's clinical state [6-8]. These conventional methods also frequently impose stringent sample inclusion criteria, which can limit their generalizability across diverse patient populations.

The widespread adoption of electronic health records (EHRs) has created unprecedented opportunities to apply advanced computational algorithms to enhance clinical decision-making. Consequently, numerous studies have leveraged EHRs data to improve sepsis-related outcomes [9-14]. Machine learning (ML) techniques, in particular, allow for the identification of latent patterns in complex patient data, offering valuable guidance for diagnosis and treatment [15-16]. Early ML applications in sepsis research often employed multivariate models like Cox

1 The First People's Hospital of Kashi, Kashi 844000, Xinjiang Uygur Autonomous Region, China

2 Southwest Petroleum University, Chengdu 610599, China

† These authors contributed equally to this work.

* Corresponding Author.

proportional-hazards and logistic regression (LR). For instance, Oami et al. [17] utilized these models to demonstrate that intensive care unit (ICU) admission significantly reduced mortality risk in sepsis patients. Similarly, Zohar et al. [18] identified hyperglycemia as an independent predictor of poor outcomes. While powerful, these statistical models are limited in their ability to handle high-dimensional, dynamic features, especially when the temporal evolution of these features is critical. To address this, methods capable of modeling temporal dynamics have gained prominence, with recurrent neural networks (RNNs) and their variants emerging as a leading approach.

Recent research has produced various sophisticated deep learning (DL) models for sepsis prediction. Shashikumar et al. [10] proposed DeepAISE, which uses a gated recurrent unit (GRU) to process time series inputs. Das et al. [19] introduced an architecture that combines an attention mechanism with a bidirectional long short term memory (Bi-LSTM) and a convolutional neural network (CNN). Other notable examples include DEAR [20], which enhances a GRU base with a two-layer attention mechanism, and SSP [21], which pairs a LSTM network with a CNN. Although these models have advanced predictive performance, they face challenges. First, they are susceptible to manual time-stamping biases in EHRs. Our analysis of the MIMIC-IV dataset identified over 160,000 laboratory records with identical specimen collection and result output times, which is inconsistent with facts. Secondly, the inherent class imbalance in sepsis prediction tasks often impacts model training and optimization. Finally, "black box" nature of many DL models hinders understanding their decision-making logic, thus impacting their clinical credibility.

To address these limitations, we propose the Feature-Wise Multi-Head Self-Attention Transformer (FW-MHSA-former) for early sepsis prediction and an Adaptive Balance-Preserving Ensemble (ABPE) training strategy. The FW-MHSA-former eschews conventional recurrent processing, instead embedding medical features as tokens and encoding temporal relationships within its attention mechanism. This design enhances robustness against the temporal inconsistencies and recording inaccuracies prevalent in EHRs data. To combat data imbalance, the ABPE method automatically partitions the majority (negative) class into multiple subsets. Each subset is then combined with the entire minority (positive) class to form a balanced sub-dataset for training an independent model instance. This ensemble approach mitigates the effects of class imbalance while preserving the statistical integrity of the original dataset. Finally, to enhance model interpretability, we visualize the attention-derived correlation matrix from the FW-MHSA-former, providing clinicians with medically plausible insights into the model's predictions.

The main contributions of this study are as follows:

- We propose the FW-MHSA-former, a novel architecture for early sepsis prediction. By mapping temporal information into its hidden dimensions, the model mitigates the impact of temporal irregularities and timestamp inaccuracies common in EHRs data while effectively extracting informative representations.
- We propose ABPE, an ensemble training strategy that constructs multiple balanced training sub-datasets. By training independent models on these datasets and aggregating their predictions via a weighted vote, this approach effectively addresses class imbalance without discarding data.
- We conduct a retrospective study on the large-scale MIMIC-IV dataset, demonstrating that our proposed method outperforms both classic ML and DL approaches in sepsis prediction, thereby validating its effectiveness and potential for clinical application.

Methods

Dataset Description

The data for this study were sourced from the Medical Information Mart for Intensive Care (MIMIC)-IV database (version 2.2) [22]. MIMIC-IV is a large-scale, publicly available, and de-identified database developed through a collaboration between the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC). It contains comprehensive clinical records from patients admitted to the ICU at BIDMC between 2008 and 2019.

Our model utilizes a curated set of features extracted from this database, including demographic data, vital signs, and an extensive panel of laboratory measurements. In addition, we incorporated several engineered features, such as interaction terms between age and blood pressure. From a physiological perspective, arterial blood pressure (ABP) is a core marker of circulatory dysfunction and a key feature of sepsis progression in the Sepsis-3 diagnostic criteria [23]. The prognostic value of blood pressure abnormalities is age-dependent: elderly patients have impaired vascular elasticity, weakened autonomic compensatory capacity, and reduced hemodynamic reserve in response to infection-induced vasodilation and circulatory collapse. Identical ABP values correspond to different sepsis risk across age groups, and this non-linear synergistic effect cannot be effectively captured by age or blood pressure alone. The inclusion of these engineered features is supported by prior research demonstrating their utility in improving sepsis prediction models [24-26]. A complete list of the features used as model inputs is provided in Table 1.

Table 1. List of demographic, vital signs, laboratory measurements, and engineered features used in the sepsis prediction model.

Category	Features
Demographic	Age
Vital signs	ABPm, ABPs, ABPD, HR, Temp, O_2, Sat, RR, FiO_2
Laboratory	WBC, INR, PT, BUN, Ven pH, Art pH, Creat, Ven CO_2, Art CO_2, Ven O_2, Art O_2 Sat, AG, Hgb ,Hct, PTT, AST, ALP, Glu, K. Mg, CL, Trop T, Fib, NH_3, HCO_3, Plt, Lactate, WB Cl, BG
Engineered	ABPm × Age, ABPs × Age, ABPd × Age, HR × Age, P/F Ratio

Definition of Sepsis Onset and Study Population

The identification of sepsis in this study followed the most recent Sepsis-3 criteria [27]. A case was labeled as sepsis when both a confirmed or suspected infection and a SOFA score increase of at least two points were observed.

In this study, each instance of ICU care was treated as a dis-

tinct sample, given that patients may have multiple ICU admissions. The samples were selected according to the following criteria:

- The patient was over the age of 18;
- The patient is not septic within the initial look-back hours following ICU admission but becomes septic at the hour of look-back + prediction time, or the patient never develops sepsis during their ICU stay;
- The number of missing records is no more than half of the look-back;
- The rate of missing values for the ABPm feature did not exceed 95%. ABPm was selected as the anchor for data quality control because it is a core indicator of circulatory function and organ perfusion, and hypotension is a key component of the Sepsis-3 definition for sepsis and septic shock. A missing rate over 95% indicates nearly no valid hemodynamic records in the observation window, which would introduce severe noise and reduce model reliability.

Compared to other studies, the inclusion criteria of this study are more tolerant of data sparsity and have no direct requirements on the frequency of feature measurement, making our work more applicable and conducive to promoting clinical application.

In the context above, the look-back refers to the time interval from which data is collected prior to the current moment, while the prediction time denotes the future interval following the current moment for which the model makes its prediction. The look-back and prediction time are both user-defined parameters that can be adjusted according to the specific requirements of the study.

Data Preprocessing

The requisite data were extracted from the MIMIC-IV dataset and subsequently adjusted to an hourly resolution by resetting minute and second fields to zero in each record. Admittedly, this hourly aggregation may smooth subtle short-term physiological variations. To further verify model robustness, we supplemented a sensitivity analysis at a 30-minute temporal resolution for comparison.

Following temporal resampling, outlier treatment was conducted. Any values less than zero were set to zero, and for each medical feature, any values exceeding three standard deviations from the mean were also set to zero. Additionally, cases involving standard deviation were handled as follows: if the standard deviation exceeded twice the mean, values greater than the mean were set to zero; if the standard deviation was between one-tenth and twice the mean, values outside five standard deviations from the mean were set to zero; and if the standard deviation was less than one-tenth of the mean, no additional processing was applied. Instead of utilizing non-original data imputation techniques, the missing data were preserved by replacing them with a value of zero. To evaluate imputation effects and justify zero-filling, we compared it against forward-filling, k-NN, and mean imputation. Using identical model settings, data splits, and metrics, we isolated the imputation strategy as the sole variable.

Negative samples spanned the full ICU stay, while positive samples included records from ICU admission up to the pre-sepsis prediction window. Only samples with a time span exceeding the sum of the look-back window and the prediction time window were retained. In the event that the records within the look-back period did not encompass the entire range of hours, missing records were populated with zeros. Ultimately, samples exhibiting over 95% missing values for ABPm were excluded. The data preprocessing steps are summarized in Figure 1.

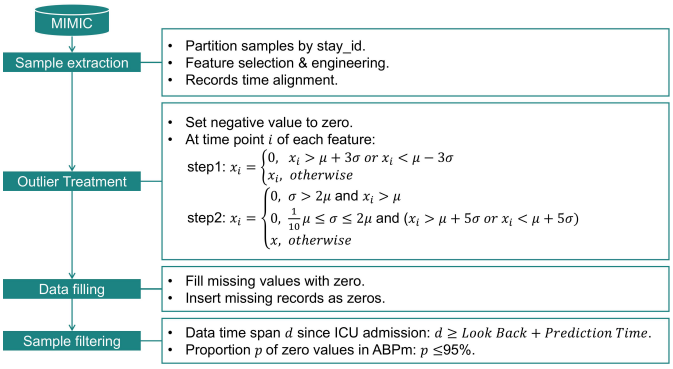


Figure 1. Data preprocessing steps for the MIMIC-IV dataset. The original data is first aligned to the hourly level, followed by outlier treatment and missing value handling.

Feature-Wise Multi-Head Self-Attention Transformer (FW-MHSA-former)

In the canonical Transformer architecture, the encoder is responsible for converting the input sequence into high-dimensional representations, while the decoder is tasked with generating the target sequence [28]. Given that sepsis prediction is a classification task rather than a sequence generation task, we have elected to employ an encoder-only Transformer architecture. The inherent manual time-stamping biases in EHRs, when combined with the alignment of records to the hour-level during preprocessing, result in the occurrence of timestamp shifts. Such discrepancies in recorded time undermine the reliability of time-step-based analysis. Therefore, In contrast to most Transformer-based models designed for time-series analysis, our approach tokenizes medical features instead of individual time points, thereby encapsulating the temporal dimension as a latent variable. This perspective enables a feature-centric representation while implicitly modeling temporal dependencies. In this configuration, each medical feature is regarded as an independent token, shifting the model's focus from cross-temporal correlations to cross-feature correlations. This mitigates the direct adverse effects of timestamp biases. Furthermore, cross-feature correlation analysis allows for the identification of interrelationships among medical features within the model's decision-making process and the assessment of each feature's influence on the model's predictions. This interpretability is intrinsic to the model and does not rely on external interpretation methods such as SHapley Additive exPlanations (SHAP), thereby providing a more direct explanatory basis for clinical decision-making.

This chapter provides a detailed introduction to the structure of the FW-MHSA-former. Figure 2 illustrates the overall structure of the proposed model.

Embedding

Embedding is a technique that maps high-dimensional or symbolic data (such as text or categorical information) into a lower-dimensional vector space [29]. This approach is widely used in fields such as Natural Language Processing (NLP)

and recommendation systems. In this study, we employ an embedding layer to map the input medical features into a continuous high-dimensional space. Specifically, the embedding layer transforms the temporal dimension, determined by the look-back size, into a higher-dimensional space with d_{model}. This transformation enables subsequent network layers to capture correlations among medical features in this enriched, high-dimensional space. In this study, the embedding operation is implemented through linear projection. Given the input data $X = x_1,\ldots,x_t \in \mathbf{R}^{(T\times N)}$, where T represents the historical observations over Y hours, and N is the number of medical features. The embedding layer maps each feature into a d_{model}-dimensional space. The transformation process can be expressed as follows:

$$H^0 = \text{Dropout}\,(XW_E + b_E) \in \mathbf{R}^{T\times d_{model}} \quad (1)$$

where $W_E \in \boldsymbol{R}^{N\times d_{model}}$ is the weight matrix of the embedding layer, $b_E \in \boldsymbol{R}^{d_{model}}$ is the bias vector, and $H^0 \in \boldsymbol{R}^{T\times d_{model}}$ is the output of the embedding layer. To prevent model overfitting, the embedded result is processed through a dropout layer, where a certain percentage of neurons is randomly dropped with a fixed probability.

Encoder

The canonical Transformer architecture employs an encoder-decoder structure. However, ongoing research has introduced both encoder-only and decoder-only architectures. The decoder-only architecture generates outputs in an autoregressive manner, rendering it well-suited for generative tasks such as text generation [30-31] and speech synthesis [32-33]. In contrast, the encoder-only architecture processes the entire input sequence in parallel, thereby enabling faster training and inference compared to decoder-only architectures. It is particularly well-suited for tasks such as classification [34-35], feature extraction [36-37], and information retrieval [38-39].

Sepsis prediction is essentially a binary classification task, where the objective is to determine whether a patient is likely to develop sepsis based on the input sequence. Consequently, we have adopted an encoder-only architecture in our study. This approach is in accordance with the latest advancements in computer vision (CV) [40-42] and time series analysis (TSA) [43-44], where encoder-only architectures have demonstrated substantial effectiveness in extracting meaningful patterns and making predictions from structured data.

The fundamental function of the encoder is to perform deep feature extraction and model the contextual relationships within the input sequence [45]. The encoder is typically composed of multiple stacked layers with identical structures. Each layer comprises two key submodules: a FW-MHSA mechanism and a Feedforward neural network (FFN). These modules are integrated via serial residual connections and layer normalization, ensuring that essential original features are retained while accelerating model convergence. The encoder's operation can be formalized as follows:

$$H'^{L} = \text{LayerNorm}\,(H^{L-1} + \text{MHSA}\,(H^{L-1})) \quad (2)$$

$$H^{L} = \text{LayerNorm}\,(H'^{L} + \text{FFN}\,(H'^{L})) \quad (3)$$

where L denotes the L-th encoder layer, and $H^L \in \boldsymbol{R}^{T\times d_{model}}$ represents the output of that layer. The FW-MHSA mechanism, represented by MHSA in Equation (2), captures global dependencies within the input sequence, while the FFN applies further nonlinear transformations to the extracted features.

Feature-Wise Multi-Head Self-Attention

The attention mechanism enables models to concentrate on the most salient parts of the input data, thereby markedly enhancing processing efficiency and performance [46]. The advent of the Transformer architecture saw the introduction of self-attention mechanisms, marking a significant shift in DL from traditional RNNs and CNNs to more efficient, parallel processing architectures. To further enhance the representational capacity of self-attention, the Transformer also introduced multi-head self-attention. This approach applies multiple attention heads in parallel, thereby improving the model's ability to capture dependencies and global information across different

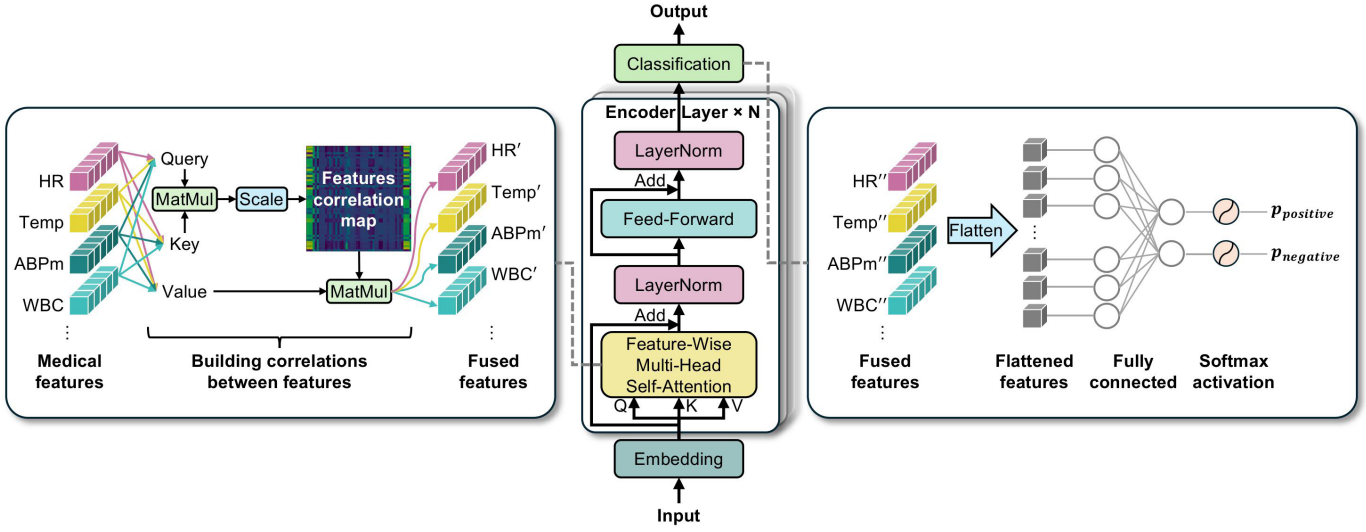


Figure 2. The architecture of the proposed FW-MHSA-former. Each encoder layer includes an input embedding layer, a Feature-Wise Multi-Head Self-Attention layer, a feed-forward neural network, and two residual connections with layer normalization. The output from the final encoder layer is fed into the fully connected layer for classification.

positions within the input sequence. In the proposed FW-MHSA-former, the multi-head self-attention mechanism is used to compute attention across input medical features in multiple feature subspaces simultaneously, allowing the model to capture more diverse dependencies between those features.

In the FW-MHSA-former, the output of the embedding layer, $H^0 \in \boldsymbol{R}^{T \times d_{model}}$, consists of T tokens, where each token represents a medical feature and is transformed as a vector in$\boldsymbol{R}^{1 \times d_{model}}$. In the feature-wise multi-head self-attention mechanism, each token is linearly transformed into a query Q, key K, and value *V* as follows:

$$Q = H^{L-1}W_Q,\ K = H^{L-1}W_K,\ V = H^{L-1}W_V \tag{4}$$

where $W_Q \in \boldsymbol{R}^{d_{model} x d_K}$, $W_K \in \boldsymbol{R}^{d_{model} x d_K}$, and $W_V \in \boldsymbol{R}^{d_{model} x d_V}$ are learnable projection matrices. Multiple sets of W_Q, W_K, andW_V are used to execute these transformations, yielding distinct Q, K, and V matrices for each attention head. For each head, the similarity between the query and key is calculated, and Softmax is applied to generate attention weights. These weights are then utilized to aggregate the value vectors, resulting in the output for the i - th head:

$$\begin{aligned} head_i &= Attention\ (QW_Q^{(i)}, KW_K^{(i)}, VW_V^{(i)}) \\ &= softmax\left(\frac{(QW_Q^{(i)}, KW_K^{(i)})^T}{\sqrt{d_k}}\right)(VW_V^{(i)}) \end{aligned} \tag{5}$$

The outputs from all heads are concatenated, and a linear transformation is applied to generate the final multi-head self-attention output:

$$MHSA(H^{L-1}) = Concat\ (head_1,...,head_h)W_O \tag{6}$$

where W_O is a projection matrix that maps the concatenated multi-head attention outputs back to the original d_{model}-dimensional space.

Classification Head

After traversing multiple encoder layers, the input sequence, represented by $H^L \in \mathbf{R}^{T \times d_{model}}$, is encoded into a high-dimensional feature representation that is rich in contextual information. The subsequent step is to transform this representation into a two-dimensional vector, which represents the probability for sepsis positive and sepsis negative. To simplify the high-dimensional feature representation, all features are first flattened into a one-dimensional vector:

$$X = Flatten(H^L) \in \boldsymbol{R}^{T \times d_{model}} \tag{7}$$

The flattening operation concatenates the features across T time steps into the vector X, thereby providing a unified input for the subsequent classification layer. This vector is then passed through a linear transformation layer, which maps it into the classification space. In this linear layer, the input features are multiplied by the weight matrix $W_{cls} \in \mathbf{R}^{(T \times d_{model}) x 2}$, thereby generating the classification output y':

$$y' = xW_{cls} + b_{cls} \in \boldsymbol{R}^2 \tag{8}$$

where $b_{cls} \in \mathbf{R}^2$ is the bias term. At this stage of the process, the model compresses the high-dimensional feature representation into the output dimension required for the classification task. To convert the output of the linear layer into a probability distribution, the Softmax function is applied:

$$\hat{y}_i = \frac{e^{y'_i}}{\sum_{j=1}^{2} e^{y'_i}},\ i = 1, 2 \tag{9}$$

The Softmax function normalizes the raw scores generated by the linear layer into a probability distribution, thereby representing the likelihood that the input sequence belongs to each class. In the context of binary classification tasks related to sepsis prediction, the output of the Softmax function compromises two probability values, each corresponding to the predicted likelihood for sepsis positive or sepsis negative.

Adaptive Balance-Preserving Ensemble (ABPE)

We propose an ABPE to avoid the issue of data imbalance in sepsis prediction. In the preliminary phase, the negative sample set is divided into multiple subsets, with each subset containing a sample size as close as possible to that of the positive sample set. Each subset of negative samples is then combined with the positive sample set to create multiple balanced sub-datasets. This even division strategy ensures the original dataset is preserved in its integrity.

The FW-MHSA-former is instantiated and trained on each sub-dataset. The area under the receiver operating characteristic curve (AUROC) score from each model serves as its respective weight. For a given prediction sample, the risk probabilities predicted by each model are weighted according to these scores. The final prediction is then derived using a weighted voting approach. The framework of the proposed ABPE and the ratio for dividing the training, validation, and test sets are illustrated in Figure 3.

Statistical Analysis

We compared baseline demographic, clinical, and laboratory characteristics between the septic and non-septic cohorts. The normality of continuous variables was evaluated using the Shapiro-Wilk test. As the data were predominantly non-normally distributed, continuous variables, such as HR and ABP, are reported as medians with interquartile ranges (IQRs), and differences between groups were assessed using the Mann-Whitney U test. Categorical variables are reported as frequencies and percentages, with inter-group comparisons performed using the Chi-squared test or Fisher's exact test, as appropriate.

To assess the prognostic performance of our FW-MHSA-former model, Kaplan-Meier (KM) survival analysis was conducted. Patients were stratified into high-risk and low-risk categories based on risk scores generated by the model. The log-rank test was then used to determine the statistical significance of any differences observed between the KM survival curves of these two groups. For all statistical tests, a two-sided $p < 0.05$ was considered the threshold for statistical significance.

Evaluation Metrics

Model performance was evaluated using accuracy, recall, the F1 score, and AUROC. Accuracy measures the overall rate of correct classifications. Recall assesses the model's ability to identify all true sepsis cases, which is critical for minimizing

missed diagnoses. The F1 score, as the harmonic mean of precision and recall, provides a balanced evaluation, especially for imbalanced datasets. Finally, the AUROC represents the model's capacity to distinguish between sepsis and non-sepsis patients. These metrics are defined as follows:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (10)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (11)$$

$$F1\ score = 2\ x\ \frac{Precision \times Recall}{Precision+Recall} \quad (12)$$

where TP (true positive) refers to the number of correctly predicted sepsis cases, TN (true negative) is the number of correctly predicted non-sepsis cases, FP (false positive) is the number of non-sepsis cases incorrectly predicted as sepsis, and FN (false negative) is the number of sepsis cases incorrectly predicted as non-sepsis.

Results

Early Sepsis Prediction and Survival Analysis

To evaluate the model's performance in the task of early sepsis prediction, we selected accuracy, recall, F1 score, and AUROC. We designed experiments with different combinations of look-back (5h, 10h, 15h, 20h) and prediction time (1h, 3h, 6h, 12h) to analyze the difference in model performance across different configurations. Table 2 presents the results of these metrics across different configurations.

Experimental results demonstrate that the optimal predictive performance across all horizons is consistently achieved with a 20-hour look-back window. Notably, the AUROC for a 1-hour lead time reached 0.94, superior to that of the 5, 10 and 15-hour configurations. A monotonic attenuation in discriminative power was observed as the prediction horizon extended; specifically, within the 20-hour window, the AUROC declined from 0.94 to 0.87 as the horizon increased from 1 to 12 hours. This phenomenon is congruent with clinical trajectories, where the stochastic nature of sepsis increases forecasting complexity over longer intervals. Furthermore, the results suggest that extended temporal depth within the look-back window provides

Table 2. The performance of the proposed method, evaluated on an independent test set for early sepsis prediction under different look-back and prediction time.

Prediction configuration		Metrics			
Look-back	Prediction time	Accuracy	Recall	F1 score	AUROC
5	1	0.86	0.83	0.85	0.93
5	3	0.83	0.78	0.82	0.89
5	6	0.79	0.72	0.77	0.87
5	12	0.76	0.70	0.74	0.83
10	1	0.83	0.82	0.83	0.91
10	3	0.83	0.81	0.82	0.90
10	6	0.81	0.76	0.80	0.86
10	12	0.75	0.72	0.74	0.82
15	1	0.86	0.88	0.86	0.92
15	3	0.83	0.82	0.83	0.91
15	6	0.79	0.74	0.78	0.88
15	12	0.79	0.77	0.78	0.86
20	1	0.87	0.90	0.87	0.94
20	3	0.86	0.84	0.86	0.93
20	6	0.79	0.84	0.86	0.93
20	12	0.80	0.83	0.80	0.87

Note: the "look-back" refers to the historical period from which patient data is used to inform the model, the "prediction time" is the lead interval for forecasting.

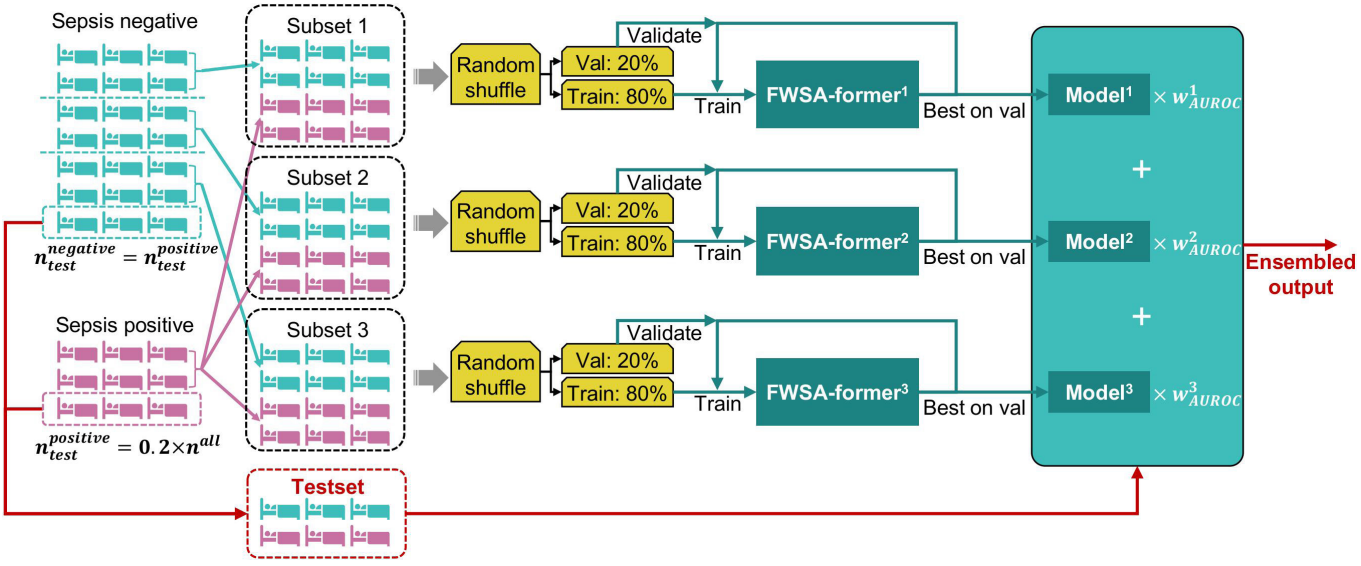


Figure 3. Overview of the proposed ABPE framework, illustrating the creation of balanced sub-datasets, model training, and the weighted voting mechanism for final prediction.

critical contextual features that augment short-term predictive stability.

To assess clinical applicability, KM survival analysis was used to stratify patients into high- and low-risk groups based on model-predicted probabilities. Figure 4 displays survival curves across various look-back and prediction windows. The results demonstrate that the model consistently distinguishes risk groups, with significantly lower survival in the high-risk cohorts. Log-rank tests confirmed all comparisons reached $p < 0.001$, indicating statistically significant survival differences.

Sensitivity Analysis of Temporal Resolution

We compared 1-hour and 30-minute resolutions to assess how temporal aggregation affects performance. Table 3 details the 12-hour prediction results by look-back window, with all other quantitative data provided in Appendix Table A.

The results show that the proposed FW-MHSA-former maintains consistent and competitive performance at the higher 30-minute temporal resolution. Across all experimental configurations, all evaluation metrics exhibited only minor fluctuations between the two resolutions, with AUROC differing by at most 0.02. This demonstrates that our model is robust to changes in time granularity and is not overly dependent on coarse hourly aggregation, further verifying the reliability and generalizability of the developed framework.

Table 3. Performance Comparison of the Proposed Model Under 1-hour and 30-minute Temporal Resolutions at a Fixed 12-hour Prediction Horizon, Stratified by Look-back Window.

Resolution	Accuracy	Recall	F1 score	AUROC
5-h look-back				
30min	0.75	0.69	0.73	0.83
1h	0.76	0.70	0.74	0.83
Δ	0.01	0.01	0.01	0.00
10-h look-back				
30min	0.74	0.70	0.73	0.80
1h	0.75	0.72	0.74	0.82
Δ	0.01	0.02	0.01	0.02
15-h look-back				
30min	0.77	0.77	0.77	0.84
1h	0.79	0.77	0.78	0.86
Δ	0.02	0.00	0.01	0.02
20-h look-back				
30min	0.78	0.78	0.78	0.85
1h	0.80	0.80	0.80	0.85
Δ	0.02	0.02	0.02	0.00

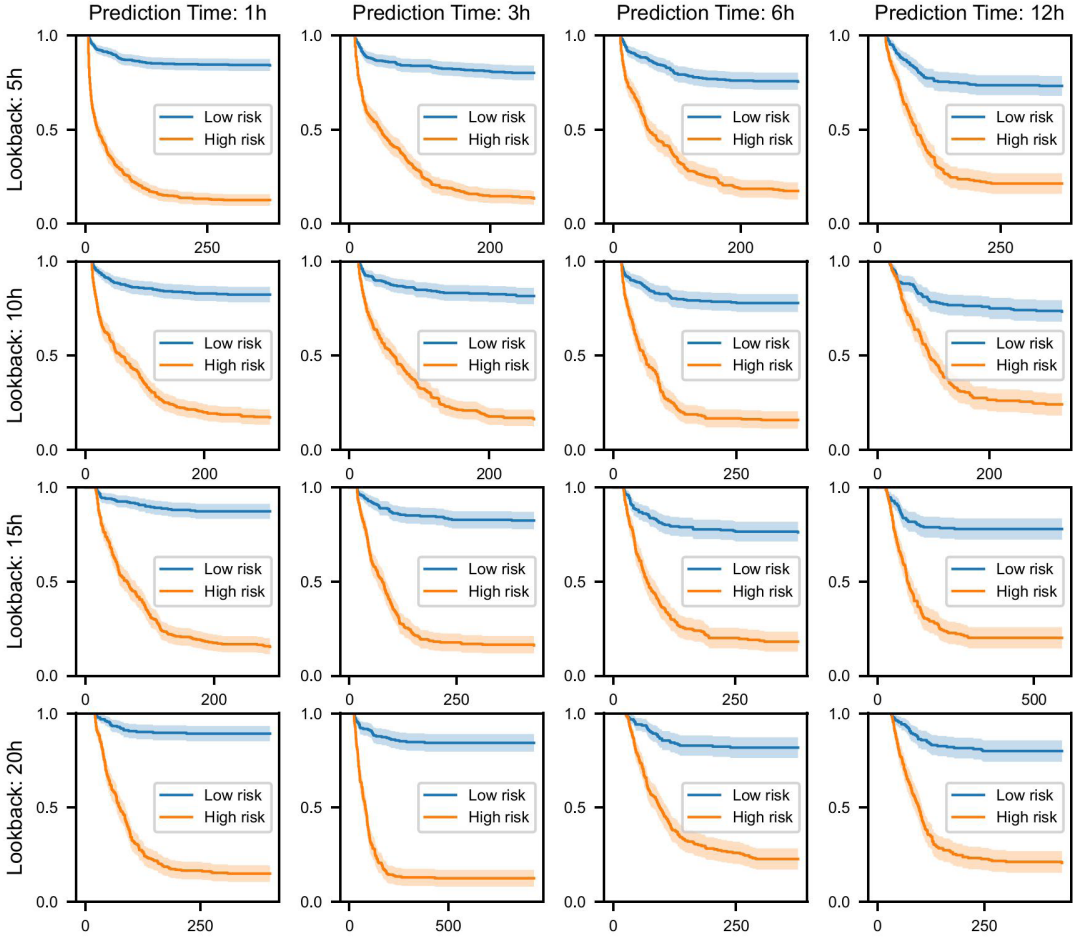


Figure 4. Kaplan-Meier survival curves comparing high-risk and low-risk sepsis groups across different look-back (5h, 10h, 15h, 20h) and prediction time (1h, 3h, 6h, 12h). The x-axis represents the survival time in hours, while the y-axis represents the survival rate. The shaded areas around each curve represent the 95% confidence intervals.

Comparison of Missing Value Imputation Strategies

Table 4 compares the performance of four missing value imputation strategies under a 20-hour look-back window; complete results for all settings are provided in Appendix Table B, C, D. Across all evaluated configurations, the zero-filling strategy consistently achieved the highest AUROC, outperforming other methods.

Notably, all four imputation methods showed consistent performance trends: model AUROC increased with the extension of the look-back window and decreased with the extension of the prediction time. This verified that the zero-filling approach used in this study does not introduce obvious artificial bias or distort the physiological distribution of vital signs and laboratory features. On the contrary, the zero-filling method better preserved the original temporal characteristics of the EHRs data, enabling the model to capture the progressive physiological deterioration trajectory of sepsis more effectively.

Comparison with classic models

Six models were selected to benchmark the proposed method. These include three classical approaches: Adaptive Boosting (AdaBoost), Long Short-Term Memory (LSTM), and Random Forest (RF). The other three are advanced attention-based or Transformer-based architectures: DuETT [47], the LSTM-Transformer hybrid, and Medformer [48]. All models were evaluated on the same test set and assessed in accordance with the evaluation metrics described in Evaluation metrics. Given that shorter look-back increases the difficulty of prediction while allowing the model to initiate sepsis prediction earlier after ICU admission, and that longer prediction times provide more response time for healthcare professionals, two representative experimental scenarios were selected for comparison. Table 5 presents the accuracy, recall, F1 score, and AUROC results for each model under a fixed 5-hour look-back with different prediction times (1h, 3h, 6h, and 12h). Table 6 illustrates the corresponding performance for distinct look-back hours (5h, 10h, 15h, and 20h) with a fixed 12-hour prediction time. The experimental results demonstrate that our method exhibits superior overall performance.

Table 4. Performance Comparison of Four Missing Value Imputation Strategies Under 20-Hour Look-Back Window Across All Prediction Time Configurations.

Method	Accuracy	Recall	F1 score	AUROC
1-h prediction				
Forward filling	0.81	0.78	0.81	0.89
K-NN imputation	0.83	0.82	0.83	0.91
Mean imputation	0.83	0.82	0.83	0.91
Zero filling	**0.87**	**0.90**	**0.87**	**0.94**
3-h prediction				
Forward filling	0.82	0.79	0.81	0.91
K-NN imputation	0.84	0.82	0.84	0.91
Mean imputation	0.82	0.81	0.82	0.91
Zero filling	**0.86**	**0.84**	**0.86**	**0.93**
6-h prediction				
Forward filling	0.74	0.76	0.74	0.81
K-NN imputation	0.75	0.74	0.75	0.81
Mean imputation	0.76	0.75	0.76	0.82
Zero filling	**0.79**	**0.83**	**0.80**	**0.87**
12-h prediction				
Forward filling	0.75	0.76	0.75	0.81
K-NN imputation	0.74	0.73	0.74	0.81
Mean imputation	0.74	0.76	0.75	0.82
Zero filling	**0.80**	**0.80**	**0.80**	**0.85**

Note: Forward filling: Missing values are filled with the last valid observation from the same patient; any remaining missing values are further imputed using the training-set mean; K-NN imputation: Missing values are imputed based on the average of the 5 nearest neighbors with complete feature profiles. Mean imputation: Missing values are filled with the mean of the corresponding feature derived from the training dataset; Zero filling: The baseline strategy in this study, which replaces missing values and out-of-range negative values with 0.

Visualization and interpretation analysis

In order to enhance the interpretability of the model, we employed the self-attention mechanism of the Transformer for sepsis prediction, encoding medical features (such as HR, ABP, WBC, etc.) as tokens and calculating correlations between these features through the attention mechanism. Figure 5 displays the correlation map from the model's final layer, obtained by averaging across all attention heads. Each cell represents the attention score when the query is sent from the corresponding row to the corresponding column.

It can be observed that ABP (including ABPm, ABPs, and ABPd) exhibit a noticeable response to almost every feature, which is reasonable given the relatively higher density of ABP data. Additionally, certain features exhibit strong uniform response patterns, including (1) ABPm, ABPs, and ABPd, (2) HR and O2 Sat, and (3) Ven CO2 and Ven O2, among others. These relationships align with known medical correlations.

While Figure 5 shows relatively low pairwise attention weights between core laboratory markers including Lactate, Creatinine, and Bilirubin, this does not indicate the model's neglect

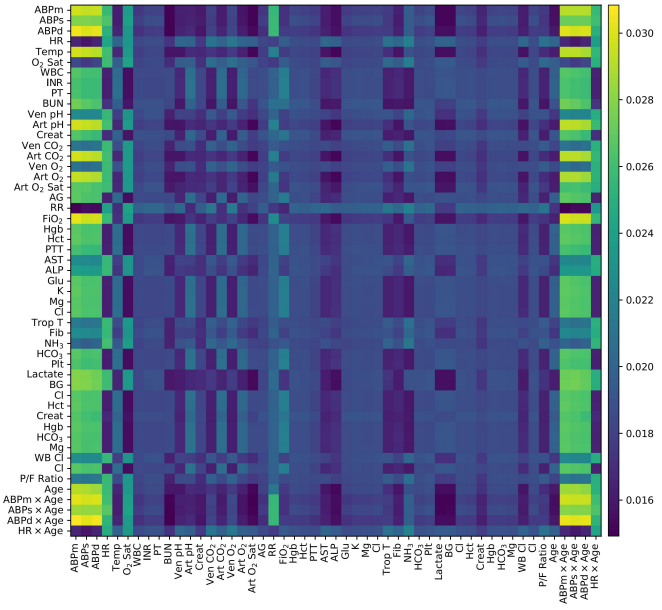


Figure 5. Correlation map of the FW-MHSA-former. Each cell represents the attention score when the query is sent from the corresponding row to the corresponding column.

Table 5. Comparison of performance metrics between our proposed method and baselines across different prediction times with a 5-hour look-back. The evaluation was conducted on the same test set, which was balanced for positive and negative samples.

Method	Accuracy	Recall	F1 score	AUROC
1-h prediction				
Adaboost	0.79	0.62	0.74	0.90
LSTM	0.79	0.64	0.76	0.90
RF	0.75	0.53	0.68	0.91
DuETT	0.85	0.81	0.82	0.92
LSTM Transformer	0.85	0.80	0.82	0.92
Medformer	0.85	0.82	0.83	**0.93**
FW-MHSA-former (Ours)	**0.86**	**0.83**	**0.85**	**0.93**
3-h prediction				
Adaboost	0.71	0.44	0.60	0.85
LSTM	0.74	0.52	0.67	0.87
RF	0.65	0.32	0.48	**0.89**
DuETT	0.81	0.75	0.81	0.88
LSTM Transformer	0.80	0.74	0.81	0.86
Medformer	0.81	0.76	0.80	0.88
FW-MHSA-former (Ours)	**0.83**	**0.78**	**0.82**	**0.89**
6-h prediction				
Adaboost	0.61	0.26	0.41	0.82
LSTM	0.63	0.29	0.44	0.80
RF	0.54	0.08	0.15	0.80
DuETT	0.78	0.69	0.72	0.85
LSTM Transformer	0.76	0.71	0.76	0.83
Medformer	0.78	0.71	0.74	0.83
FW-MHSA-former (Ours)	**0.79**	**0.72**	**0.77**	**0.87**
12-h prediction				
Adaboost	0.54	0.11	0.19	0.77
LSTM	-	-	-	0.67
RF	0.50	0.00	0.01	0.75
DuETT	0.75	0.63	0.68	**0.83**
LSTM Transformer	0.74	0.67	0.71	0.82
Medformer	0.74	0.69	0.73	0.82
FW-MHSA-former (Ours)	**0.76**	**0.70**	**0.74**	**0.83**

Note: The "-" denotes undefined values due to no positive predictions, which can occur when the model fails to identify any minority class instances.

of organ dysfunction— a key point of the Sepsis-3 definition centered on organ dysfunction [49]. Instead, this phenomenon is attributed to the sparse and irregular measurement of these laboratory markers in clinical practice, which leads to the lack of strong static correlations between them. Importantly, the model does not rely solely on high-frequency vital signs; rather, it effectively captures the dynamic temporal changes of these core laboratory markers that are associated with organ failure, and integrates them with vital signs and engineered features (e.g., P/F Ratio, ABP×Age) that directly reflect organ function. This integration fully aligns with the Sepsis-3 definition [49], confirming that the model has not overlooked the core multi-organ failure logic of sepsis.

The correlation map illustrates the relationships between medical features, with each column representing the attention scores when responding to queries from other features. The higher the attention score, the more information from the queried feature is incorporated into the output of the querying feature. The more a feature is incorporated by other features, the greater its influence on the model's final output. By summing the attention scores in each column of the correlation map, we calculated an importance score, indicating the contribution of each medical feature to the model's predictions. These importance scores were ranked in descending order, as shown in Figure 6.

Table 6. Comparison of performance metrics between our proposed method and baselines across different prediction times with a 12-hour prediction. The evaluation was conducted on the same test set, which was balanced for positive and negative samples.

Method	Accuracy	Recall	F1 score	AUROC
20-h look-back				
Adaboost	0.63	0.31	0.46	0.82
LSTM	0.64	0.31	0.46	0.76
RF	0.50	0.01	0.01	0.82
DuETT	0.78	0.75	0.77	0.84
LSTM Transformer	0.79	0.78	0.79	**0.85**
Medformer	0.78	0.75	0.77	0.83
FW-MHSA-former (Ours)	**0.80**	**0.80**	**0.80**	**0.85**
15-h look-back				
Adaboost	0.58	0.21	0.33	0.80
LSTM	-	-	-	0.66
RF	0.51	0.01	0.22	0.77
DuETT	0.75	0.73	0.75	**0.86**
LSTM Transformer	0.77	0.75	0.77	0.85
Medformer	0.76	0.71	0.76	0.84
FW-MHSA-former (Ours)	**0.79**	**0.77**	**0.78**	**0.86**
10-h look-back				
Adaboost	0.56	0.15	0.25	0.79
LSTM	-	-	-	0.55
RF	-	-	-	0.79
DuETT	0.74	0.71	0.73	0.81
LSTM Transformer	0.72	0.69	0.71	0.80
Medformer	0.72	0.70	0.72	0.80
FW-MHSA-former (Ours)	**0.75**	**0.72**	**0.74**	**0.82**
5-h look-back				
Adaboost	0.54	0.11	0.19	0.77
LSTM	-	-	-	0.67
RF	0.50	0.00	0.01	0.75
DuETT	0.75	0.69	0.73	0.82
LSTM Transformer	0.75	0.65	0.71	0.82
Medformer	**0.76**	0.68	0.70	**0.83**
FW-MHSA-former (Ours)	**0.76**	**0.70**	**0.74**	**0.83**

Note: The "-" denotes undefined values due to no positive predictions, which can occur when the model fails to identify any minority class instances.

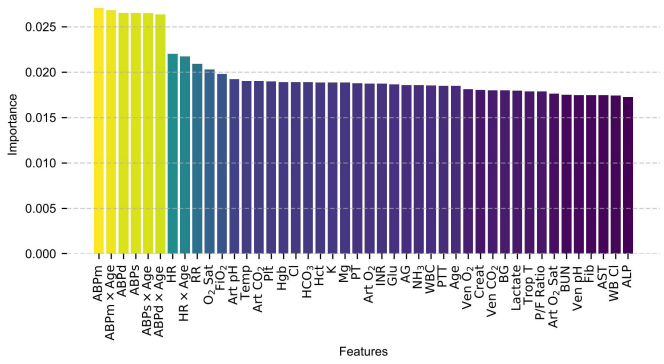


Figure 6. Feature importance ranking based on the sum of attention scores in the correlation map. The y-axis represents the medical feature importance score, while the x-axis displays the corresponding feature names.

Discussions

This study introduces and validates a novel DL framework that synergizes a FW-MHSA-former with an ABPE strategy. The architecture was specifically engineered to overcome two fundamental challenges inherent in EHRs data: the management of temporally irregular observations and the profound class imbalance typical of medical datasets. Evaluated retrospectively on the comprehensive, publicly accessible MIMIC-IV dataset, the model demonstrated robust performance. Across a range of configurations, varying the look-back window from 5 to 20 hours and the prediction horizon from 1 to 12 hours, the model consistently yielded an AUROC exceeding 0.80. It achieved a peak AUROC of 0.94 when utilizing a 20-hour look-back window for a 1-hour prediction horizon. Furthermore, in comparative analyses, our framework outperformed six conventional ML and DL models.

A detailed investigation into the model's performance characteristics revealed crucial temporal dynamics. Experimental results demonstrate that the model's predictive performance scales with the extension of the look-back window, peaking at 20 hours, although performance remains comparable between the 15-hour and 20-hour windows specifically at a 12-hour lead time. This finding suggests that physiological data within the 20-hour window preceding sepsis onset encapsulates the most critical information for risk prediction, as longer temporal sequences better capture the complete trajectory of progressive clinical deterioration. Such insight provides an empirical basis for determining optimal data acquisition windows in clinical settings. Furthermore, while the observed decay in performance as the lead time extends aligns with the established principle that predictive uncertainty increases with temporal distance, the model maintains robust discriminative capacity (AUROC > 0.82) even at a 12-hour lead time. This underscores its potential to establish a clinically meaningful window for early intervention.

An analysis of the model's interpretability confirmed that it learned patterns highly congruent with established clinical pathophysiology. Feature importance analysis revealed that the highest-weighted predictors were engineered interaction terms, namely ABPm×Age, ABPd×Age, and ABPs×Age. This finding is significant, as it indicates that the prognostic value of blood pressure in sepsis is non-linear and substantially modulated by age. This interaction term more effectively captures the heightened hemodynamic vulnerability of older patients during infectious stress than individual metrics alone—a nuance frequently missed by traditional linear models. Following these were foundational vital signs such as HR, RR, Temp, and O2 Sat, which are the cornerstones of rapid screening tools like the qSOFA. This alignment demonstrates that the model independently learned decision-making criteria consistent with expert clinical consensus. Moreover, laboratory markers reflecting organ dysfunction and tissue hypoperfusion—including Plt, Creat, and lactate—also received high importance scores, which is consistent with the Sepsis-3 definition centered on the SOFA score. Thus, the model functions not as an inscrutable 'black box' but as a system whose decision-making logic is largely transparent and capable of augmenting our understanding of sepsis pathophysiology.

While our model demonstrated exemplary performance in a retrospective context, its ultimate clinical value is contingent upon prospective validation. Kaplan-Meier survival analysis confirmed the model's ability to stratify patients into high- and low-risk cohorts with statistically significant differences in survival outcomes ($p < 0.001$), laying a critical foundation for its translation into clinical decision support tools. Notably, with a stable AUROC of 0.82–0.86 across all look-back configurations, the 12-hour prediction window fits well into routine ICU workflows based on 12-hour nursing shift cycles, providing reliable evidence to assist real-time clinical scheduling and hierarchical patient management. This lead time aligns perfectly with routine patient assessment and handover procedures: the model can be run once per shift at handover to generate a 12-hour sepsis risk forecast for the entire on-duty period, enabling proactive rather than reactive care. For patients stratified as high-risk, this 12-hour lead time enables concrete, guideline-aligned interventions: upgrading from 4-hourly routine vital sign monitoring to 1-hourly continuous hemodynamic monitoring, pre-emptive completion of blood culture and inflammatory biomarker testing, and advance preparation of vascular access and fluid resuscitation protocols. This fully addresses the core clinical requirement of the Surviving Sepsis Campaign guidelines, which mandate initiation of evidence-based care within 1 hour of sepsis recognition, by providing a sufficient window for risk mitigation before clinical deterioration. Furthermore, the model's robust risk stratification enables optimized ICU resource allocation: low-risk patients (with consistently negative predictions) can avoid unnecessary invasive monitoring and frequent laboratory testing, while clinical resources are prioritized to the high-risk cohort identified by the model.

Nevertheless, this study is subject to several limitations. First, we standardized the time series by aggregating raw records to a uniform hourly resolution, which may smooth high-frequency physiological fluctuations and subtle acute physiological changes. To mitigate this limitation and evaluate the influence of temporal granularity, we supplemented a sensitivity analysis using a 30-minute interval resolution. Experimental results demonstrated that our model yields stable and comparable performance at both 1-hour and 30-minute temporal resolutions, confirming that the prediction capability is not substantially compromised by hourly aggregation and supporting the robustness of the proposed method.

Second, the model was developed and validated exclusively using data from the MIMIC-IV database. Although large and of

high quality, MIMIC-IV represents a single academic medical center. Therefore, the model's performance and generalizability must be confirmed through external validation on datasets from different healthcare systems and patient populations.

Conclusions

This study introduces a novel DL architecture that integrates a FW-MHSA-former with an ABPE for the early prediction of sepsis. The proposed framework exhibits superior predictive performance, enabling the robust stratification of patients into high- and low-risk cohorts, thereby enhancing clinical decision support and optimizing patient management strategies. A critical finding is the model's exceptional robustness and reliability, as evidenced by its sustained high-fidelity predictive accuracy across diverse and challenging look-back and forecast time configurations. These results underscore the framework's significant clinical utility and its potential to improve patient outcomes through more timely and precise.

Abbreviations

electronic health records: EHRs; deep learning: DL; Feature-Wise Multi-Head Self-Attention Transformer: FW-MHSA-former; Adaptive Balance-Preserving Ensemble: ABPE; Medical Information Mart for Intensive Care IV: MIMIC-IV; Kaplan-Meier: KM; Area Under the Receiver Operating Characteristic Curve: AUROC; machine learning: ML; Systemic Inflammatory Response Syndrome: SIRS; Sequential Organ Failure Assessment: SOFA; quick SOFA: qSOFA; intensive care unit: ICU; logistic regression: LR; recurrent neural network: RNNs; gated recurrent unit: GRU; bidirectional long short term memory: Bi-LSTM; convolutional neural network: CNN; Massachusetts Institute of Technology: MIT; Beth Israel Deaconess Medical Center: BIDMC; SHapley Additive Explanations: SHAP; natural language processing: NLP; computer vision: CV; time series analysis: TSA; Feedforward neural network: FFN; interquartile ranges: IQRs; true positive: TP; true negative: TN; false positive: FP; false negative: FN; Adaptive Boosting: AdaBoost; Random Forest: RF; Support Vector Classifier: SVC.

Author Contributions

Mireaye Abudurexiti: Conceptualization, Investigation, Supervision, Writing – Original Draft. Jianshu Wang: Methodology, Visualization, Validation, Software. Pengfei Zhang: Data Curation, Investigation, Supervision. Fangfang Liu: Resources, Project Administration. Zhiqiang Jia: Funding Acquisition, Supervision, Investigation, Writing – Review and Editing.

Acknowledgments

Not Applicable.

Funding Information

This study was supported by the Xinjiang Uygur Autonomous Region Health Science and Technology Program (2025001CG-ZHYDXM653124504), the "Tianshan Talent" High-Level Medical and Health Personnel Training Program (TSYC202401B223) and Kashgar Prefecture Science and Technology Plan Program (KS2024022).

Ethics Approval and Consent to Participate

This study uses data from the MIMIC-IV database, which consists of anonymized records that do not require individual patient consent.

Competing Interests

The authors declare no conflicts of interest relevant to this article.

Data Availability

The MIMIC-IV dataset utilized in this study originates from the Beth Israel Deaconess Medical Center (BIDMC) and is managed by the Massachusetts Institute of Technology's Computer Science and Artificial Intelligence Laboratory (MIT CSAIL). Given that the dataset contains sensitive patient health information, researchers must apply for access and complete a Data Use Agreement (DUA) along with Collaborative Institutional Training Initiative (CITI) certification to ensure ethical usage and protection of privacy. Consequently, the raw data is not publicly available; however, authorized researchers may obtain access through the official MIMIC-IV portal at https://mimic.physionet.org/.

References

[1] Duan, Y., Huo, J., Chen, M., Hou, F., Yan, G., Li, S., et al. (2023). Early prediction of sepsis using double fusion of deep features and handcrafted features. Applied intelligence (Dordrecht, Netherlands), 1–17. Advance online publication. https://doi.org/10.1007/s10489-022-04425-z

[2] Fleuren, L. M., Klausch, T. L. T., Zwager, C. L., Schoonmade, L. J., Guo, T., Roggeveen, L. F., et al. (2020). Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive care medicine, 46(3), 383–400. https://doi.org/10.1007/s00134-019-05872-y

[3] Reyna, M. A., Josef, C. S., Jeter, R., Shashikumar, S. P., Westover, M. B., Nemati, S., et al. (2020). Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. Critical care medicine, 48(2), 210–217. https://doi.org/10.1097/CCM.0000000000004145

[4] Lauritsen, S. M., Kalør, M. E., Kongsgaard, E. L., Lauritsen, K. M., Jørgensen, M. J., Lange, J.,et al. (2020). Early detection of sepsis utilizing deep learning on electronic health record event sequences. Artificial intelligence in medicine, 104, 101820. https://doi.org/10.1016/

j.artmed.2020.101820
[5] Wang, Y., Lu, X., Tan, J., He, J., Yin, H., Chen, P., et al. (2025). The cGAS-STING Pathway: Insights into Regulatory Mechanisms, Disease Dysregulation, and Therapeutic Development. Life Conflux, 2(1), e300. https://doi.org/10.71321/dr57c347
[6] Bomrah, S., Uddin, M., Upadhyay, U., Komorowski, M., Priya, J., Dhar, E., et al. (2024). A scoping review of machine learning for sepsis prediction- feature engineering strategies and model performance: a step towards explainability. Critical care (London, England), 28(1), 180. https://doi.org/10.1186/s13054-024-04948-6
[7] Deng, H. F., Sun, M. W., Wang, Y., Zeng, J., Yuan, T., Li, T., et al. (2021). Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. iScience, 25(1), 103651. https://doi.org/10.1016/j.isci.2021.103651
[8] Delahanty, R. J., Alvarez, J., Flynn, L. M., Sherwin, R. L., & Jones, S. S. (2019). Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis. Annals of emergency medicine, 73(4), 334–344. https://doi.org/10.1016/j.annemergmed.2018.11.036
[9] Yan, M. Y., Gustad, L. T., & Nytrø, Ø. (2022). Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. Journal of the American Medical Informatics Association: JAMIA, 29(3), 559–575. https://doi.org/10.1093/jamia/ocab236
[10] Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Critical care medicine, 46(4), 547–553. https://doi.org/10.1097/CCM.0000000000002936
[11] Shashikumar, S. P., Josef, C. S., Sharma, A., & Nemati, S. (2021). DeepAISE - An interpretable and recurrent neural survival model for early prediction of sepsis. Artificial intelligence in medicine, 113, 102036. https://doi.org/10.1016/j.artmed.2021.102036
[12] Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., et al. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications, 11(1), 3852. https://doi.org/10.1038/s41467-020-17431-x
[13] Mao, Q., Jay, M., Hoffman, J. L., Calvert, J., Barton, C., Shimabukuro, D., et al. (2018). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ open, 8(1), e017833. https://doi.org/10.1136/bmjopen-2017-017833
[14] Tang, X., Zheng, D., Kebede, G. S., Li, Z., Li, X., Lu, C., et al. (2023). An automatic segmentation framework of quasi-periodic time series through graph structure. Applied Intelligence, 53(20), 23482–23499. https://doi.org/10.1007/s10489-023-04814-y
[15] van de Sande, D., van Genderen, M. E., Huiskens, J., Gommers, D., & van Bommel, J. (2021). Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive care medicine, 47(7), 750–760. https://doi.org/10.1007/s00134-021-06446-7
[16] Lai, H., Wu, G., Zhong, Y., Chen, G., Zhang, W., Shi, S., et al. (2023). Red blood cell distribution width improves the prediction of 28-day mortality for patients with sepsis-induced acute kidney injury: A retrospective analysis from MIMIC-IV database using propensity score matching. Journal of intensive medicine, 3(3), 275–282. https://doi.org/10.1016/j.jointm.2023.02.005
[17] Oami, T., Imaeda, T., Nakada, T. A., Abe, T., Takahashi, N., Yamao, Y., et al. (2023). Mortality analysis among sepsis patients in and out of intensive care units using the Japanese nationwide medical claims database: a study by the Japan Sepsis Alliance study group. Journal of intensive care, 11(1), 2. https://doi.org/10.1186/s40560-023-00650-x
[18] Zohar, Y., Zilberman Itskovich, S., Koren, S., Zaidenstein, R., Marchaim, D., & Koren, R. (2021). The association of diabetes and hyperglycemia with sepsis outcomes: a population-based cohort analysis. Internal and emergency medicine, 16(3), 719–728. https://doi.org/10.1007/s11739-020-02507-9
[19] Das, P.P., Wiese, L., Mast, M., Böhnke, J., Wulff, A., Marschollek, M., et al. (2024). An attention-based bidirectional LSTM-CNN architecture for the early prediction of sepsis. International Journal of Data Science and Analytics, 20, 1841 - 1855. https://doi.org/10.1007/s41060-024-00568-z
[20] Zhao, Y., Wu, Y., Liu, M., Cai, X., Zhang, Y., Yuan, X. (2022). DEAR: Dual-Level Self-attention GRU for Online Early Prediction of Sepsis. In: Zhao, X., Yang, S., Wang, X., Li, J. (eds) Web Information Systems and Applications. WISA 2022. Lecture Notes in Computer Science, vol 13579. Springer, Cham. https://doi.org/10.1007/978-3-031-20309-1_37
[21] Rafiei, A., Rezaee, A., Hajati, F., Gheisari, S., & Golzan, M. (2021). SSP: Early prediction of sepsis using fully connected LSTM-CNN model. Computers in biology and medicine, 128, 104110. https://doi.org/10.1016/j.compbiomed.2020.104110
[22] Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific data, 10(1), 1. https://doi.org/10.1038/s41597-022-01899-x
[23] Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., et al. (2016). Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 762–774. https://doi.org/10.1001/jama.2016.0288
[24] Seymour, C. W., Kennedy, J. N., Wang, S., Chang, C. H., Elliott, C. F., Xu, Z., et al. (2019). Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA, 321(20), 2003–2017. https://doi.org/10.1001/jama.2019.5791
[25] Gyawali, B., Ramakrishna, K., & Dhamoon, A. S. (2019). Sepsis: The evolution in definition, pathophysiology, and management. SAGE open medicine, 7, 2050312119835043. https://doi.org/10.1177/2050312119835043
[26] Camacho-Cogollo, J. E., Bonet, I., Gil, B., & Iadanza, E. (2022). Machine Learning Models for Early Prediction of Sepsis on Large Healthcare Datasets. Electronics, 11(9), 1507. https://doi.org/10.3390/electronics11091507
[27] Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The Third

International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 801–810. https://doi.org/10.1001/jama.2016.0287
[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762
[29] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv, 1301.3781. https://arxiv.org/abs/1301.3781
[30] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. arXiv, 2005.14165. https://arxiv.org/abs/2005.14165
[31] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI.
[32] Wang, X., Thakker, M., Chen, Z., Kanda, N., Eskimez, S.E., Chen, S., et al. (2023). SpeechX: Neural Codec Language Model as a Versatile Speech Transformer. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, 3355-3364. https://doi.org/10.1109/TASLP.2024.3419418
[33] Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., et al. (2023). UniAudio: An Audio Foundation Model Toward Universal Audio Generation. ArXiv, 2310.00704. https://arxiv.org/abs/2310.00704
[34] Chen, C.-F., Fan, Q., & Panda, R. (2021). CrossViT: Cross-attention multi-scale vision transformer for image classification. arXiv, 2103.14899. https://arxiv.org/abs/2103.14899
[35] Xin, C., Liu, Z., Zhao, K., Miao, L., Ma, Y., Zhu, X., et al. (2022). An improved transformer network for skin cancer classification. Computers in biology and medicine, 149, 105939. https://doi.org/10.1016/j.compbiomed.2022.105939
[36] Dutta, P., Sathi, K. A., Hossain, M. A., & Dewan, M. A. A. (2023). Conv-ViT: A Convolution and Vision Transformer-Based Hybrid Feature Extraction Method for Retinal Disease Detection. Journal of imaging, 9(7), 140. https://doi.org/10.3390/jimaging9070140
[37] Wang, L., Fang, S., Meng, X., & Li, R. (2022). Building extraction with vision transformer. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–11. https://doi.org/10.1109/TGRS.2022.3186634
[38] Bhopale, A. P., & Tiwari, A. (2024). Transformer based contextual text representation framework for intelligent information retrieval. Expert Systems with Applications, 238, 121629. https://doi.org/10.1016/j.eswa.2023.121629
[39] Kim, Y., Bang, S., Sohn, J., & Kim, H. (2022). Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers. Automation in Construction, 134, 104061. https://doi.org/10.1016/j.autcon.2021.104061
[40] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv, 2010.11929. https://arxiv.org/abs/2010.11929
[41] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992–10002. https://doi.org/10.48550/arXiv.2103.14030
[42] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., et al. (2022). Swin transformer V2: Scaling up capacity and resolution. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11999–12009. https://doi.org/10.1109/CVPR52688.2022.01170
[43] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. arXiv. https://doi.org/10.48550/arXiv.2211.14730
[44] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., et al. (2023). iTransformer: Inverted transformers are effective for time series forecasting. arXiv. https://doi.org/10.48550/arXiv.2310.06625
[45] Mo, Y., Wu, Q., Li, X., & Huang, B. (2021). Remaining useful life estimation via transformer encoder enhanced by a gated convolutional unit. J Intell Manuf 32, 1997–2006. https://doi.org/10.1007/s10845-021-01750-x
[46] Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., et al. (2022). Attention mechanisms in computer vision: A survey. Computational Visual Media, 8(3), 331–368. https://doi.org/10.1007/s41095-022-0271-y
[47] Labach, A., Pokhrel, A., Huang, X., Zuberi, S., Yi, S., Volkovs, M., et al. (2023). DuETT: Dual Event Time Transformer for Electronic Health Records. arXiv, 2304.13017. https://doi.org/10.48550/arXiv.2304.13017
[48] Wang, Y., Huang, N., Li, T., Yan, Y., Zhang, X. (2024). Medformer: A Multi-Granularity Patching Transformer for Medical Time-Series Classification. arXiv, 2405.19363. https://doi.org/10.48550/arXiv.2405.19363
[49] Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 801–810. https://doi.org/10.1001/jama.2016.0287