Research Square

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

DTKF: A Sepsis Early Prediction Framework Based on Large Language Model Semantic Augmentation and Dual-Source Knowledge Distillation

Yitian Zhang
Shanghai University of Engineering Science
Xinlei Ji
University of Shanghai for Science and Technology
Huijie Zhang
Shanghai University of Engineering Science
Liang Zhou
Shanghai University of Medicine and Health Sciences
Jiachen Guo
Shanghai University of Medicine and Health Sciences
Lei Yin
yinl@sit.edu.cn

Shanghai Institute of Technology

Research Article

Keywords: Sepsis Early Prediction, Semantic Feature Augmentation, Dual-Source Knowledge Distillation, Large Language Models, Multi-modal Fusion

Posted Date: April 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9199143/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 October 12th, 2026. See the published version at https://doi.org/10.1186/s12911-026-03870-x.

DTKF: A Sepsis Early Prediction Framework Based on Large Language Model Semantic Augmentation and Dual-Source Knowledge Distillation

Yitian Zhang[1],Xinlei Ji[2],Huijie Zhang[3]

[1,3]Shanghai University of Engineering Science, Shanghai,China

[2]University of Shanghai for Science and Technology,Shanghai,China

[*]Corresponding Author:Liang Zhou,Jiachen Guo,Lei Yin(yinl@sit.edu.cn)

Abstract

Background Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection. Early and accurate prediction is crucial for reducing mortality. However, clinical settings often face challenges such as data scarcity, class imbalance, and the underutilization of unstructured clinical notes. Existing scoring systems and traditional machine learning models primarily rely on structured vital signs, failing to capture the rich semantic risk factors embedded in textual data. Although Large Language Models (LLMs) excel in text understanding, their high computational costs and potential hallucinations hinder direct deployment in resource-constrained emergency settings.

Methods To address these limitations, this study proposes a novel Dual-Teacher Knowledge Fusion (DTKF) framework. First, we introduce a semantic feature augmentation module that leverages DeepSeek to extract semantic risk probabilities from clinical notes and explicitly fuses them with structured vital signs to construct

high-order hybrid features for the student model. Second, to prevent overfitting under small-sample conditions, we design a dual-source knowledge distillation strategy. This strategy integrates rule-based knowledge and data-driven semantic knowledge to construct hybrid teacher signals, regulating a lightweight logistic regression student model via soft-label supervision.

Results Experiments on the MIMIC-IV dataset demonstrate that DTKF achieves an AUC of 0.796. This result not only significantly outperforms traditional unimodal baselines but also improves performance by 6.6% compared to direct prediction using LLMs.

Conclusions This study effectively integrates rule-based and data-driven knowledge, enhances feature representation using textual information, and improves model performance and robustness through distillation techniques, providing a clinically valuable solution for sepsis early warning.

Keywords Sepsis Early Prediction, Semantic Feature Augmentation, Dual-Source Knowledge Distillation, Large Language Models, Multi-modal Fusion.

Background

Sepsis is a life-threatening organ dysfunction syndrome where early and precise prediction is critical for reducing patient mortality[1,2]. While the Sepsis-3 consensus emphasizes organ dysfunction assessment based on the SOFA score, early warning scoring systems remain pivotal for rapid bedside screening in Emergency Departments (ED) and general wards[3]. Among various systems, the National Early Warning Score

(NEWS) is widely recommended due to its standardized physiological parameters and superior clinical utility[4]. Evidence suggests that NEWS outperforms qSOFA in sensitivity and accuracy for early sepsis detection in ED settings[5]. The core advantages of such rule-based systems lie in their simplicity, stability, and high interpretability, making them ideal for resource-constrained clinical environments.

However, these systems face significant limitations. They rely primarily on limited structured vital signs, often failing to meet the precision required for early warning[6]. More importantly, they cannot effectively incorporate unstructured clinical notes, which contain latent patterns and semantic cues crucial for early identification. To address these gaps, researchers have extensively explored machine learning[7,8]. While theoretically capable of modeling complex non-linear relationships, these data-driven methods encounter practical hurdles in the ED. First, high-performance models typically demand extensive structured data, which is often sparse in early triage stages[9]. Second, the extreme scarcity of positive sepsis cases creates severe class imbalance, causing algorithms to be biased towards the majority class and increasing the risk of missed diagnosis[10], a challenge that persists even with resampling techniques[11]. Crucially, traditional ML methods struggle to deeply parse the semantics of medical text[12,13], and the difficulty in fusing textual features with structured parameters results in the loss of vital warning signals[14].

Recently, LLMs like DeepSeek have revolutionized the extraction of semantic information from clinical text[15]. Their pre-trained knowledge and reasoning capabilities enable the extraction of precise risk indicators from unstructured notes[16].

Nevertheless, deploying standalone LLMs in resource-constrained environments is hindered by high computational costs, inference latency, and potential hallucinations[17]. Consequently, a promising direction lies in synergizing the powerful semantic understanding of LLMs with the stability of clinical scores (NEWS) and the efficiency of lightweight models to effectively tackle data scarcity and imbalance.

Knowledge Distillation (KD) offers a paradigm for transferring knowledge from complex to lightweight models[18,19]. We innovatively extend this framework by constructing a dual-source teacher that fuses rule-based clinical knowledge (i.e.,NEWS) with data-driven semantic insights (i.e.,LLMs). NEWS provides physiological stability, while the LLMs capture semantic sensitivity; together, they form complementary knowledge sources. Furthermore, we employ KD as a regularization strategy to transfer the teacher's generalization ability to a lightweight student model, mitigating overfitting issues caused by extreme data imbalance[20].

Methods

This study proposes a sepsis early prediction framework based on semantic augmentation and dual-source knowledge distillation, named DTKF. The framework aims to break the performance bottleneck of traditional unimodal models by integrating structured vital signs and unstructured clinical notes to achieve precise quantification of sepsis risk. As shown in Figure 1, DTKF consists of three core modules: dual-teacher construction, semantic feature augmentation, and knowledge transfer optimization.

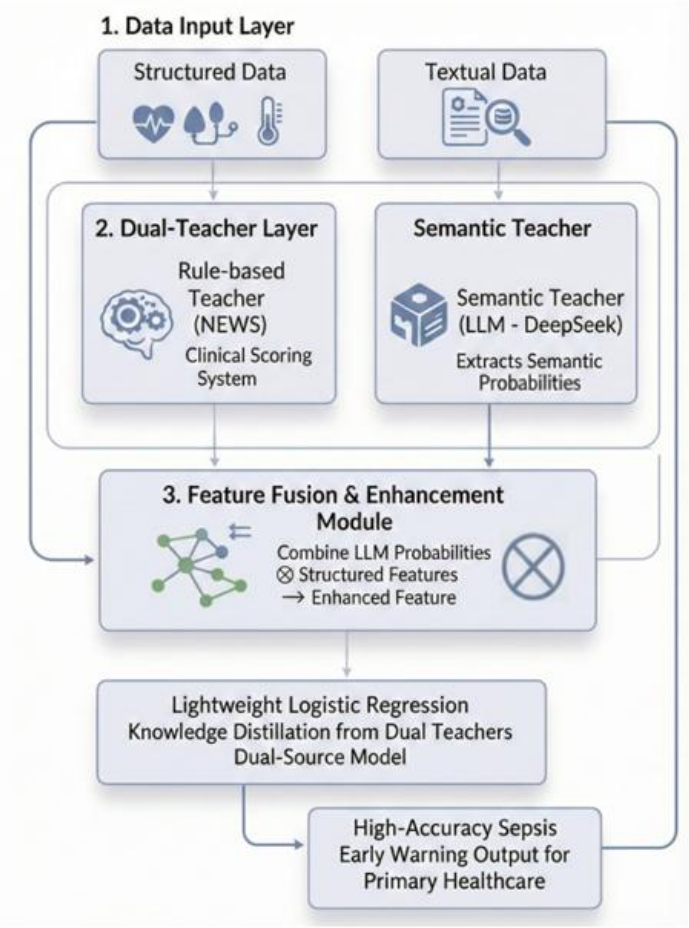


Fig. 1. Schematic overview of the proposed Dual-Teacher Knowledge Fusion framework.

Problem Definition and Framework Overview

Notation Definition

Consider a sepsis prediction dataset $D=\{(x_i, t_i, y_i)\}_{i=1}^{N}$ containing N samples, where $y_i \in \{0,1\}$ represents the true binary label (0 for negative, 1 for sepsis positive). Each sample contains two parts of heterogeneous data:

(1) Structured Data($x_i \in \mathbb{R}^d$): Contains $d = 7$ basic vital signs and laboratory indicators. Specifically: Consciousness, Systolic Blood Pressure (SBP), Heart Rate (HR), Respiratory Rate (RR), Temperature (Temp), Oxygen Saturation (SpO2), and

White Blood Cell count (WBC). These indicators follow the NEWS scoring standards and are the most easily accessible parameters in ED settings.

(2) Unstructured Text(t_i): Contains ED triage records, chief complaints, and preliminary diagnosis notes, containing rich contextual semantic information.

The goal of this study is to construct a lightweight mapping function $f_\theta: (x_i, t_i) \rightarrow \hat{y}_i$that enables the model to fully utilize cross-modal information for precise prediction while maintaining low computational complexity.

DTKF Framework Process

Addressing the challenge of heterogeneous and unevenly distributed clinical data modalities, the DTKF framework adopts a dual-stream parallel processing architecture (as shown in Figure 1):

(1) Data Stream A (Rule Channel): Structured data x_i is fed into the NEWS scoring system to calculate a clinical rule-based risk score, which is further mapped to a rule probability P_{rule}.

(2) Data Stream B (Semantic Channel): Unstructured text t_i is fed into the DeepSeek Large Language Model. Through Chain-of-Thought (COT) reasoning, the model extracts a latent semantic risk probability P_{sem}.

In the feature fusion stage, we explicitly fuse P_{sem} as a high-order feature with x_i to generate an enhanced feature vector x_{aug}, which serves as the input for the lightweight logistic regression. During the training phase, the student model is optimized under the soft-label supervision of dual-source hybrid teacher signals ($P_{rule}$and P_{sem}) to achieve effective knowledge transfer.

Dual-Source Teacher Construction

To transform heterogeneous clinical prior knowledge into soft labels usable for distillation, we construct specific teacher signals for structured data and unstructured text, respectively.

Rule-Based Teacher: NEWS Score Probability Space Mapping

The original NEWS score ($g \in \mathbb{Z}, 0 \leq g \leq 20$) consists of discrete integers, which poses a challenge for directly serving as continuous probability supervision in knowledge distillation. To bridge the gap between discrete medical rules and computable risk probabilities, we employ the Normal Cumulative Distribution Function (NCDF) for non-linear mapping.

Considering the right-skewed distribution of NEWS scores and the non-linear progression of physiological deterioration, we calibrated the mapping parameters to μ=7 and σ=4. Specifically, the mean μ=7 aligns with the established clinical consensus where a NEWS score ≥7 signifies a critical threshold for "high clinical risk." The standard deviation σ=4 is selected to ensure the probability curve smoothly covers the distribution range of the vast majority of samples, thereby sensitively capturing risk gradients in higher score segments. The mapping formula is defined as:

$$P_{NEWS_CDF=}\Phi(\frac{g-7}{4})(\ g\in Z,\ 0\leq g\leq 20) \quad (1)$$

Through this transformation, the rigid, discrete scoring rule is reconstructed into a continuous confidence signal anchored by clinical logic, providing high-quality supervision for the subsequent knowledge distillation process.

Semantic Teacher: LLM-based Risk Quantification via Chain-of-Thought

For unstructured text, we utilized the DeepSeek to construct a CoT reasoning process containing three stages to extract the denoised semantic risk probability P_{sem}; as shown in Figure 2.

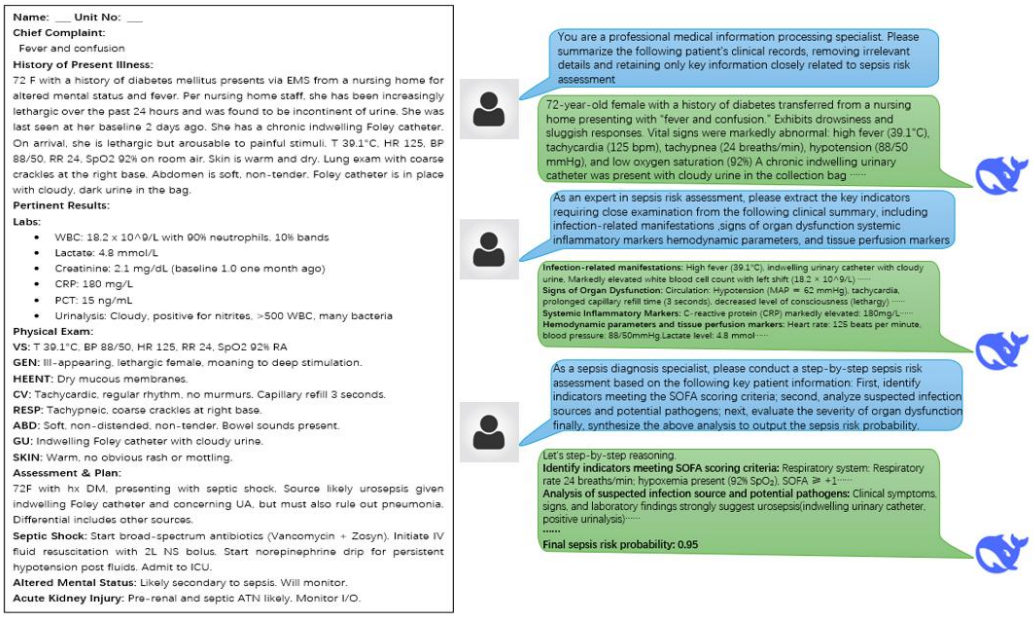


Fig. 2. Pipeline for generating the LLM-based sepsis prediction probability (P_LLM), involving prompt-based summarization, key feature extraction, and CoT reasoning.

First, addressing the common issues of redundant information and noise in clinical text, this study constructs a prompt template to implement text summarization and denoising. The template is designed as: "You are a professional medical information processing expert. Please summarize the following patient clinical records, removing irrelevant details and retaining key information closely related to sepsis risk assessment." This step significantly compresses text length and eliminates irrelevant noise while preserving core clinical indicators related to sepsis.

Second, based on the compressed text, multi-dimensional features closely related

to sepsis risk are further extracted. The prompt template is designed as: "As a sepsis risk assessment expert, please extract key indicators for focus from the following clinical summary, including: infection-related manifestations, organ dysfunction signs, systemic inflammation markers, hemodynamic parameters, and tissue perfusion markers." This step outputs a structured set of high-dimensional features, providing a semantic basis for subsequent probabilistic reasoning.

Finally, a CoT reasoning mechanism is introduced to perform multi-step logical reasoning and probability generation. Instead of a direct query, the model is prompted to act as a sepsis diagnostic expert and follow a structured diagnostic pathway. Specifically, it is instructed to evaluate the extracted features by identifying SOFA-related indicators, analyzing potential infection sources and pathogens, and assessing organ dysfunction severity. Ultimately, the model synthesizes these sequential evaluations to output a final sepsis risk probability. This method enhances the reliability and clinical acceptability of prediction results through an interpretable reasoning path.

Semantic Feature Enhancement Module

Traditional logistic regression models rely solely on linear feature combinations, making it difficult to capture complex non-linear relationships between "physiological indicators" and "textual risk." To this end, we designed a semantic feature enhancement module to explicitly construct high-order fusion features.

Let the standardized original structured feature vector be $v_{raw} \in \mathbb{R}^d$. The input vector $x_{student}$ for the student model is constructed as follows:

$$x_{student} = \text{Concat}(v_{raw}, P_{sem}, v_{inter}) \in \mathbb{R}^{2d+1} \quad (2)$$

Where v_{inter} is the semantic-physiological synergistic feature, defined as the element-wise product (Hadamard Product) of the semantic probability and physiological indicators:

$$v_{inter} = P_{sem} \otimes v_{raw} = [P_{sem} * x^{(1)}, P_{sem} * x^{(2)}, \dots, P_{sem} * x^{(d)}] \quad (3)$$

Introducing the interaction term essentially endows the model with dynamic gating capability. When the LLMs deem the textual risk to be extremely low, the interaction term approaches 0, and the model's decision degrades to relying mainly on original signs. When the LLMs identifie high-risk semantics accompanied by abnormal physiological indicators, the interaction term is significantly activated, forming a "text + numerical" double confirmation mechanism. This design significantly improves the lightweight model's feature representation capability for complex cases without increasing network depth.

Knowledge Transfer and Optimization

To achieve robust prediction in scenarios with small samples and class imbalance, we define the student model as a lightweight logistic regression:

$$\hat{y} = \sigma(w^T x_{student} + b) \quad (4)$$

During training, we use a hybrid loss function, simultaneously minimizing the deviation from true labels and the difference from dual-source teacher distributions:

$$L_{total} = \alpha L_{CE} + (1 - \alpha) L_{KD} \quad (5)$$

Where L_{CE} is the standard binary cross-entropy loss, used to supervise the learning of true labels. L_{KD} is the knowledge distillation loss, using Kullback-Leibler (KL) divergence to measure the difference between student prediction and teacher

distribution:

$$L_{KD} = T^2 \cdot KL(P_{teacher} \,||\, P_{student}) (6)$$

To maximize the complementarity of dual-source knowledge, we discarded traditional linear λ weighting and instead constructed a structured asymmetric teacher distribution. Specifically, given the LLMs sensitivity in capturing early symptoms, we designate it as the guide for the positive class; given the specificity of the NEWS score in physiological stability, we designate it as the guide for the negative class.

We first construct the unnormalized dual-source probability vector $\tilde{P}$:

$$\tilde{P}_{pos} = P_{sem} \text{(Derived from LLM)}$$

$$\tilde{P}_{neg} = 1 - P_{rule} \text{(Derived from NEWS)}$$

Since these two probabilities come from heterogeneous spaces and do not sum to 1, we introduce a normalization operation to generate the final teacher soft label $P_{teacher}$:

$$P_{teacher}^{(k)} = \frac{\tilde{p}_k}{\tilde{p}_{pos} + \tilde{p}_{neg}}, k \in \{pos, neg\} (7)$$

Through this normalization mechanism, the model can dynamically adjust the direction of gradients based on the relative confidence of the two modalities, allowing the semantic model and the rule model to take over their respective specialized risk intervals without manually setting a hyperparameter λ.

In experiments, the remaining temperature parameter T and balance coefficient α are hyperparameters optimized via grid search on the validation set. This dual supervision mechanism ensures that the model can both fit the data distribution and internalize dual prior knowledge from rules and semantics.

Results

To construct a student model input that is lightweight and easy to obtain, we selected 7 basic physiological and laboratory indicators. The reasons for selecting these variables are: First, they are standard parameters most easily obtained in emergency triage; second, they are highly consistent with the components of the NEWS scoring system, facilitating the alignment of rule knowledge. Specific features include: Consciousness status, SBP, HR, RR, Temp, SpO2, and WBC.

Baseline characteristics of the screened patients are shown in Table 1. Statistical test results show that, except for temperature, the sepsis group and the non-sepsis group exhibit significant statistical differences on all included physiological indicators ($p<0.001$), verifying the effectiveness of the selected features in distinguishing sepsis risk.

Table 1 Baseline Characteristics of the MIMIC-IV dataset.

Sepsis_label	**0(N=10117)**	**1(N=4333)**	**p values**
Consc(Status)			<0.001
A	9525(94.1%)	3703(85.4%)	
B	505(4.9%)	497(11.4%)	
C	87(0.8%)	133(3.0%)	
SBP(mmHg)	121.4(20.96)	116.5(22.91)	<0.001
HR(beats/min)	82.4(16.83)	88.0(19.99)	<0.001
RR(breaths/min)	19.1(5.23)	20.2(5.89)	<0.001
SpO2(%)	96.3(3.52)	96.6(3.49)	<0.001
Temp (℃)	36.9(1.16)	36.9(1.52)	0.0490
WBC($\times 10^9/L$)	10.5(8.53)	12.8(11.85)	<0.001
NEWS Score	5.23(1.97)	5.96(2.23)	<0.001
llm_prob*HR	29.3(20.63)	54.3(27.47)	<0.001
llm_prob*WBC	3.82(5.22)	8.13(9.96)	<0.001
llm_prob*NEWS Score	1.90(1.60)	3.74(2.31)	<0.001

In this binary classification study, we label sepsis patients as the positive class (label 1) and the remaining patients as the negative class (label 0). Considering the common issues of limited data and class imbalance in ED scenarios, this study focuses on two aspects: (1) Model performance based on knowledge transfer when training sample size is small; (2) Robustness of the model when facing class imbalance. To this end, we designed the following experimental protocol to systematically evaluate the method's performance under different training set sizes and different class distributions.

First, by adjusting the ratio of class samples in the raw data, we constructed simulated datasets with varying degrees of imbalance. Specifically, six negative-to-positive sample ratios were established: 10:1, 1:1, 5:1, 10:1, 50:1, and 100:1. The corresponding actual sample sizes (Negative, Positive) were set as (2000, 200), (2000, 2000), (10,000, 2000), (10,000, 1000), (10,000, 200), and (10,000, 100), respectively. These configurations served as the overall sample pool, which was subsequently partitioned into training, validation, and test subsets. Second, the overall dataset was divided according to different proportions. Training set sizes were set to 5%, 10%, 20%, and 40%, respectively. The validation set was fixed at 20% of the total, and the remainder was used as the test set. Stratified sampling was used during division to maintain the original class distribution.

Finally, to improve result credibility and statistical stability, we performed 100 resampling iterations for the above process to obtain the distribution of AUC values, thereby reflecting the model's generalization ability more comprehensively and robustly.

Comparative Experimental Results Analysis

Table 2 Performance comparison of the proposed method against direct LLM-based prediction baselines on the MIMIC-IV dataset.

Binary results(AUC)	
Models	**MIMIC-IV dataset**
LLM[1]	0.7423
LLM[2]	0.7474
LLM_KLD[1]	0.7499
LLM_KLD[2]	0.7538
DTKF	**0.7955**

LLM[1]: Direct prediction using the large language model (LLM) with raw structured and textual data, without any prompt engineering.

LLM[2]: LLM prediction enhanced with COT prompting.

LLM_KLD[1/2]: Corresponding LLM configurations further refined via knowledge distillation.

DTKF: Our proposed framework integrating NEWS clinical scores and LLM semantic probabilities.

To systematically evaluate the potential and limitations of LLMs in this task and verify the synergistic effect of CoT prompts and knowledge distillation, this study first designed comparative experiments with LLMs. As shown in Table 2, structured patient data was first converted into natural language descriptions and input into the DeepSeek model alongside original clinical text. On this basis, we compared the effects of two prompting strategies: LLM[1] represents basic prompting (direct prediction), while LLM[2] introduces CoT prompting.Experimental results show that the introduction of the CoT strategy effectively improved LLM prediction performance,

with the AUC of LLM[2] achieving a notable increase compared to LLM[1]. This indicates that structured logical reasoning can moderately enhance the LLM's semantic understanding and judgment capabilities in complex medical decision tasks. Secondly, knowledge distillation technology further optimized LLM prediction performance; the distilled models saw AUC improvements to 0.750 and 0.754 respectively.

However, although both CoT and independent knowledge distillation brought gains, the performance ceiling of these single-source knowledge models still lags significantly behind the DTKF framework proposed in this study. Relying solely on the LLM's text semantic knowledge, despite using perfect prompt engineering or distillation techniques, still struggles to break through inherent bottlenecks. Organic fusion of the LLM's data-driven knowledge with the clinical rule knowledge contained in the NEWS score is the key to achieving a leap in model performance.

Table 3 Performance comparison of different models under varying training set ratios and class imbalance ratios on the MIMIC dataset.

Binary results(AUC)							
Baseline Models	**R1** / **P1**	**100:1**	**50:1**	**10:1**	**5:1**	**1:1**	**1:10**
NEWS	0.05	0.619	0.628	0.632	0.630	0.634	0.650
LR		0.556	0.580	0.648	0.658	0.652	0.627
LR_KLD		0.593	0.623	0.659	0.665	0.664	0.663
DTKF		**0.734**	**0.756**	**0.759**	**0.765**	**0.765**	**0.741**
NEWS	0.1	0.610	0.627	0.633	0.631	0.636	0.649
LR		0.558	0.617	0.659	0.665	0.665	0.662
LR_KLD		0.612	0.612	0.655	0.665	0.667	0.656
DTKF		**0.729**	**0.741**	**0.763**	**0.765**	**0.768**	**0.750**
NEWS	0.2	0.625	0.628	0.634	0.630	0.634	0.643
LR		0.602	0.648	0.667	0.670	0.671	0.676
LR_KLD		0.611	0.630	0.668	0.670	0.670	0.656
DTKF		**0.746**	**0.741**	**0.763**	**0.768**	**0.767**	**0.759**

NEWS		0.611	0.617	0.631	0.630	0.634	0.645
LR	0.4	0.643	0.652	0.668	0.673	0.680	0.686
LR_KLD		0.641	0.629	0.661	0.669	0.681	0.660
DTKF		**0.756**	**0.763**	**0.764**	**0.765**	**0.773**	**0.765**

P1 represents the proportion of samples in the training set to simulate scenarios with scarce clinical samples, while R1 represents the ratio of positive to negative cases among sepsis patients to simulate scenarios with imbalanced clinical samples.

Furthermore, to verify the comprehensive performance against baseline models under different sample ratios, as shown in Table 3, in the most challenging scenario (extreme data scarcity and imbalance: Training Ratio P1=0.05, Imbalance Ratio R1=100:1), the proposed DTKF framework still demonstrated exceptional sepsis early warning capability, reaching an AUC of 0.734, significantly outperforming all baselines. As the training data ratio increased and class distribution improved, model performance further increased, reaching an optimal AUC of 0.773 (at P1=0.4, R1=1:1). Even in the most challenging extreme data scenarios, the performance gap with the optimal result was relatively limited. In contrast, the traditional NEWS system and Logistic Regression model showed limited performance improvement and larger fluctuations across different training sizes, performing particularly unstably under high imbalance. This fully validates the effectiveness and robustness of the multi-source knowledge fusion and distillation mechanism in dealing with data scarcity and extreme class imbalance.

Additionally, across different class ratio settings, DTKF consistently maintained optimal performance, with AUC fluctuation ranges far smaller than other methods, demonstrating excellent robustness. Particularly under the double challenge of extreme

imbalance and training sample scarcity, the framework maintained high discriminatory ability, indicating its potential for deployment in real-world clinical environments with limited resources.

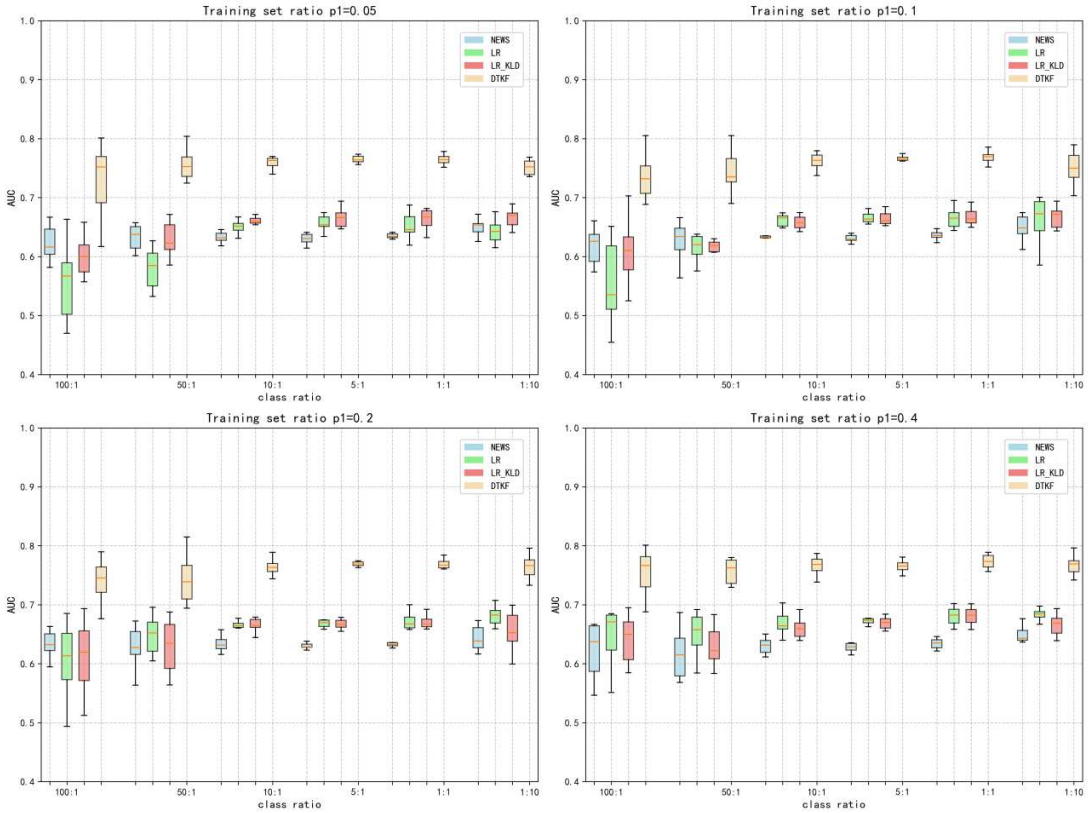


Fig. 3. Boxplots of classification performance for multiple methods under varying training set ratios and class ratios.

According to the boxplots in Figure 3, the DTKF method fusing NEWS clinical priors and LLM reasoning priors consistently exhibits the characteristics of "highest box position, narrowest interquartile range, and tightest whisker range." This indicates that this method not only significantly outperforms NEWS, traditional LR, and KD models based only on NEWS in average AUC, but also shows smaller performance fluctuations across multiple experimental iterations. This reflects stronger performance stability; the multi-source prior knowledge distillation mechanism can more stably capture key patterns for sepsis early identification, reducing the interference of random

factors on model performance.

As the class ratio evolves from balanced to extremely imbalanced, the boxplots for NEWS, traditional LR, and LR_KLD show a clear downward trend in position and a sharp expansion in IQR and whisker range. This indicates that discrimination performance decays significantly and stability drops sharply for these methods as imbalance intensifies, due to the domination of the majority class and insufficient learning of minority class features.

In contrast, although the DTKF boxplot shows slight fluctuations in extremely imbalanced scenarios, the box remains in a significantly higher AUC range than other methods, and the expansion of the IQR/whisker range is far smaller. This verifies the core advantage of the Dual-Source Asymmetric Teacher Distribution: by fusing NEWS clinical domain priors with LLM general reasoning priors to provide more discriminatory soft-label guidance, it effectively mitigates the constraint of data imbalance on minority class learning.

When the training set ratio p1 increases from 0.05 to 0.4, the DTKF box remains in the high AUC zone, and the IQR narrows further as p1 increases. Sufficient training data further strengthens the transfer efficiency of multi-source priors, further improving stability. While NEWS, LR, and LR_KLD show some improvement with increasing p1, overall AUC remains lower, and fluctuations are consistently larger than DTKF. This further demonstrates that DTKF has stronger robustness to training data volume: it compensates for data deficiency with multi-source priors when data is scarce, and fully fuses data and prior information to achieve dual optimization of performance and

stability when data is sufficient.

Ablation Study Analysis

To explore the contribution of multi-source features and the role of knowledge transfer components, we conducted feature ablation and KD component ablation experiments in the extreme imbalance scenario (Training p1=0.05, Sample Ratio 100:1). Results are shown in Figure 4.

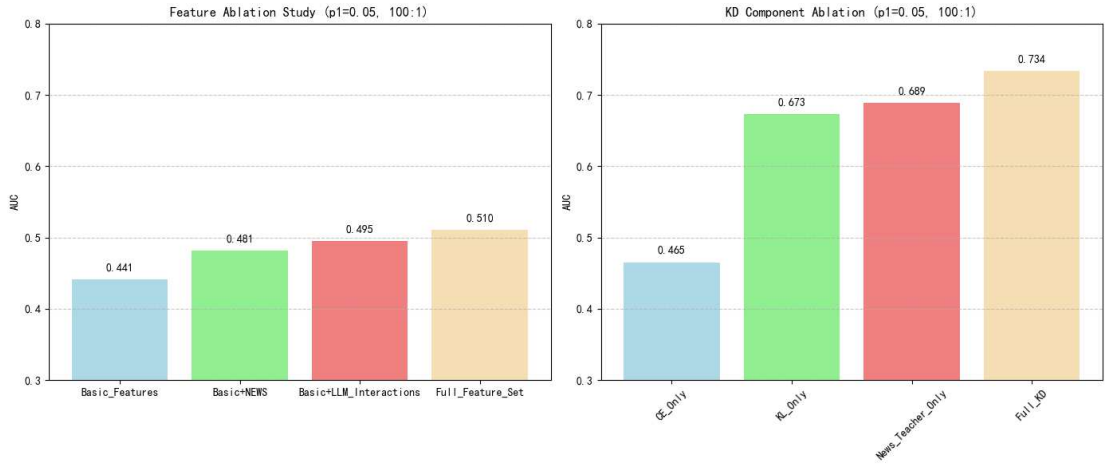


Fig. 4. Performance comparison of feature ablation and knowledge distillation component ablation experiments. The left section illustrates model performance under different feature combinations, Basic_Features includes only basic physiological parameters; Basic+NEWS combines basic features with the NEWS score; Basic+LLM_Interactions combines basic features with large language model interaction features and Full_Feature_Set represents the complete set of all features. The right section displays performance under different knowledge distillation configurations, including: using only cross-entropy loss (CE_Only), using only KL divergence loss (KL_Only), using only the NEWS score as the teacher signal (News_Teacher_Only), and the full knowledge distillation method (Full_KD).

Feature Ablation Experiment

Feature ablation aims to evaluate the individual and combined contributions of basic physiological features, NEWS score, and LLM fused features using a logistic

regression model without knowledge transfer. When using only basic physiological features, the model AUC is only 0.441, indicating that single physiological parameters have limited discriminatory ability for early sepsis, especially under extreme imbalance and scarcity. After adding the NEWS total score derived from vital signs as an extra feature, the AUC improved.

Further adding the fusion feature of basic signs and the LLM yielded an AUC of 0.495. This reflects that the fusion term of LLM prediction and physiological features can mine the association pattern between "model reasoning priors and physiological indicators," injecting additional information dimensions for sepsis identification.

When using the Full Feature Set, AUC reached 0.510, the highest among all feature combinations. This verifies the complementarity of multi-source features: basic physiological features provide "direct observational information," NEWS provides "clinical assessment priors," and LLM fusion features integrate "model reasoning priors." The synergy of the three more comprehensively characterizes the complex pathological state of early sepsis. However, the AUC remains low (just above 0.5), confirming that Logistic Regression performs poorly in predicting under extreme imbalance and minimal sample sizes without distillation.

Knowledge Transfer Component Ablation Experiment

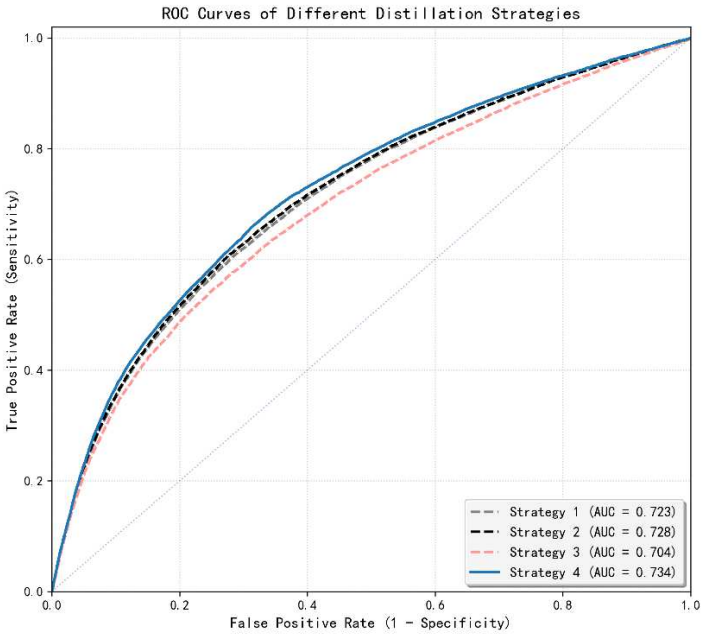


Figure 5. Comparison of ROC curves for different knowledge distillation strategies in early sepsis prediction.The curves are plotted using pooled prediction results from 100 independent iterations to ensure statistical robustness.

Strategy 1: Baseline symmetric distillation with fixed equal weights (λ=0.5)

Strategy 2: Symmetric distillation with optimal weights determined via grid search.

Strategy 3: Original asymmetric distillation, where the LLM guides the negative class and NEWS guides the positive class.

Strategy 4: The proposed Semantics-Dominant Asymmetric Distillation, where the LLM guides the positive class (capturing early semantic symptoms) and NEWS guides the negative class (serving as a specificity filter).

The KD component ablation analyzed the impact of "Loss Function Type" and "Teacher Signal Source." When using only Cross-Entropy (CE) loss, AUC was only 0.465. In the 100:1 imbalance scenario, supervised learning is easily dominated by the majority class. Using only KL divergence loss raised AUC to 0.673. This shows that

KD can transfer information via the teacher's "soft label" probability distribution, alleviating learning bias caused by imbalance.

When the teacher signal came only from NEWS, AUC further improved to 0.689. This reflects that NEWS, as a "clinical domain prior," provides targeted guidance. Under the Full_KD configuration (Strategy 4), the model AUC reached 0.734, the highest among all configurations.

The core reasons for this result lie in the synergy of two aspects:

(1)Teacher Signal Level: We constructed a Semantics-Dominant Asymmetric Teacher Distribution. Results from Figure 5 (Strategy 4) show that using LLM as the Positive Class Teacher and NEWS as the Negative Class Teacher is best. This reveals a profound clinical logic: unstructured information in clinical text often precedes the deterioration of vital signs. Therefore, the LLM can capture earlier risk signals, dominating positive prediction; conversely, the NEWS score is based on objective signs—when NEWS is low, sepsis risk is excluded with high probability, making it suitable for dominating negative prediction. This complementary mechanism corrects the traditional NEWS score's defect of "insufficient sensitivity" in early warning.

(2)Loss Function Level: CE loss ensures fitting accuracy to "true labels," while KL divergence transfers "uncertainty information" from the teacher distribution. Their synergy enables the student model to learn both "intrinsic data patterns" and "prior knowledge," ultimately achieving optimal performance in extreme imbalance scenarios.

Discussion

Principal Findings and Interpretations

This study presents the DTKF framework, a novel sepsis early prediction model that synergizes clinical rule-based knowledge with data-driven semantic insights via dual-source knowledge distillation. Our experiments on the MIMIC-IV dataset demonstrate that DTKF achieves an AUC of 0.796, significantly outperforming traditional unimodal baselines and achieving comparable performance to direct LLM prediction while maintaining the efficiency of a lightweight model. The core innovation lies in the successful fusion of structured vital signs (NEWS) and unstructured clinical notes (LLM-extracted semantics) to address the persistent challenges of data scarcity and extreme class imbalance in emergency settings.

Comparison with Clinical Scoring Systems and Traditional ML

Our findings address critical gaps identified in existing clinical scoring systems and machine learning approaches. Traditional rule-based systems like NEWS and qSOFA are valued for their stability and interpretability[4, 5]. However, as noted in previous studies, these systems rely heavily on a limited set of structured physiological parameters, leading to a plateau in discriminatory performance and an inability to capture latent semantic cues from clinical text [6, 14]. Our results confirm this limitation; the standalone NEWS score in our experiments showed lower sensitivity compared to the fusion model.

To overcome these barriers, prior researchers have applied machine learning

algorithms such as Random Forest, XGBoost, and stacking frameworks[21-23]. While these methods generally outperform simple scoring systems by modeling non-linear relationships, they often struggle with the extreme class imbalance and data sparsity characteristic of early triage stages[9-11]. As shown in our comparative analysis (Table 3), traditional Logistic Regression and NEWS failed to maintain stability under extreme imbalance (100:1 ratio). In contrast, DTKF utilized the dual-teacher signals to effectively regularize the student model, demonstrating superior robustness and generalization in small-sample scenarios.

Synergy of LLMs and Knowledge Distillation

The integration of Large Language Models represents a paradigm shift in medical text mining[29]. Recent studies have demonstrated the efficacy of LLMs and CoT reasoning in extracting precise risk indicators from unstructured notes[15, 27, 28]. However, the direct deployment of models like DeepSeek or ChatGPT in resource-constrained environments is hindered by high inference latency and computational costs[17]. Our study validates these concerns; while direct LLM prediction yielded high accuracy (Table 2), it is impractical for real-time bedside screening.

By employing KD, we bridged the gap between the high semantic capability of LLMs and the efficiency requirements of clinical practice. While KD has been previously explored in clinical time-series[30, 31], and specifically for sepsis prediction using NEWS as a teacher, these approaches typically relied on a single knowledge source. Ji et al.[32] utilized NEWS to guide a student model but missed the rich semantic context provided by textual data. Our Dual-Teacher framework advances this by

combining the physiological "specificity" of NEWS (dominating the negative class guidance) with the semantic "sensitivity" of the LLM (dominating the positive class guidance). This asymmetric distillation strategy ensures that the lightweight student model inherits both clinical stability and semantic reasoning capabilities, achieving the "best of both worlds."

Limitations and Future Directions

Despite significant results, limitations exist. First, the framework relies on external LLMs to process clinical text during the teacher construction phase, which, although done offline, implies a dependence on the quality of the LLM's pre-training. Second, the current model uses a static snapshot of data and has not yet integrated dynamic time-series modeling capabilities to capture the temporal evolution of sepsis, a direction explored by other deep learning studies [24-26]. Third, while the logistic regression student has parameter interpretability, the correspondence between its decision logic and clinical pathways requires further analysis and external validation.

Future work will focus on: (1) exploring locally deployed lightweight LLMs to further reduce dependencies and enhance data security; (2) introducing temporal modeling mechanisms to handle irregular sampling in clinical data; and (3) conducting multi-center prospective clinical validation to further evaluate generalization ability and clinical utility.

Conclusion

Sepsis early prediction is a key challenge in critical care medicine. This study

proposed a Dual-Source Knowledge Fusion (DTKF) framework based on knowledge distillation. It successfully integrates clinical rule knowledge and semantic information extracted from unstructured text by Large Language Models. Through knowledge distillation, the fused knowledge is transferred to a lightweight logistic regression student model, achieving a synergistic improvement in performance and interpretability. Experimental results on the MIMIC-IV dataset show that the DTKF framework significantly outperforms baseline methods in AUC, robustness, and stability, particularly in scenarios with extreme data scarcity and class imbalance.

List of abbreviations

DTKF: Dual-Teacher Knowledge Fusion

LLM: Large Language Model

ED: Emergency Department

NEWS: National Early Warning Score

qSOFA: quick Sequential Organ Failure Assessment

ML: Machine Learning

KD: Knowledge Distillation

CDF: Cumulative Distribution Function

KL: Kullback-Leibler

CE: Cross-Entropy

LR:Logistic Regression

AUC: Area Under the Curve

MIMIC: Medical Information Mart for Intensive Care

CoT: Chain-of-Thought

SOFA: Sequential Organ Failure Assessment

SBP: Systolic Blood Pressure

HR: Heart Rate

RR: Respiratory Rate

WBC: White Blood Cell count

SpO_2:Oxygen Saturation

Temp: Temperature

IQR: Interquartile Range

ROC:Receiver Operating Characteristic

Declarations

Ethical approval and consent to participate

This study utilized de-identified data from the publicly available Medical Information Mart for Intensive Care (MIMIC-IV) database (version v3.1) and its associated clinical notes module (MIMIC-IV-Note, version 2.2). The use of the MIMIC database was approved by the Institutional Review Boards (IRBs) of Beth Israel Deaconess Medical Center (BIDMC) and the Massachusetts Institute of Technology (MIT). As the data are de-identified and publicly released, the requirement for informed consent was waived. All methods were carried out in accordance with relevant guidelines and regulations, including the data use agreement for the MIMIC database.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests in this section.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

Yitian Zhang was responsible for conceptualization, literature review, methodology, writing the original draft, and visualization and figures. Xinlei Ji was responsible for investigation, data curation, formal analysis, and validation. Huijie Zhang contributed to literature review, data collection, and assisted with the initial draft preparation. Lei Yin was responsible for conceptualization, supervision, writing—review and editing, as well as providing academic support and critical revision. Liang Zhou, Jiachen Guo, and Lei Yin served as corresponding authors, jointly responsible for the overall direction and organization of the research, overall supervision, and ensuring the integrity and accuracy of the work.

Acknowledgments

The authors would like to thank the developers and maintainers of the Medical Information Mart for Intensive Care (MIMIC) database for providing access to the clinical data used in this study. We also acknowledge the use of Google Scholar for literature retrieval and preliminary reference collection during the research process.

Reference

1. Singer, M.; Deutschman, C.S.; Seymour, C.W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G.R.; Chiche, J.-D.;Coopersmith, C.M.; et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016, 315,801–810

2. Evans, L.; Rhodes, A.; Alhazzani, W.; Antonelli, M.; Coopersmith, C.M.; French, C.; Machado, F.R.; Mcintyre, L.; Ostermann, M.;Prescott, H.C.; et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. Crit.Care Med. 2021, 49, e1063–e1143

3. Bone RC, Balk RA, Cerra FB, et al. Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis. Chest 1992;101:1644-55.

4. McGinley A, Pearse R M. A national early warning score for acutely ill patients[J]. Bmj, 2012, 345.

5. Durr D, Niemi T, Despraz J, et al. National Early Warning Score (NEWS) outperforms quick sepsis-related organ failure (qSOFA) score for early detection of sepsis in the emergency department[J]. Antibiotics, 2022, 11(11): 1518.

6. C. Wang, R. Xu, Y. Zeng, Y. Zhao, X. Hu, A comparison of qSOFA, SIRS and NEWS in predicting the accuracy of mortality in patients with suspected sepsis: a metaanalysis, PLoS One 17 (2022) e0266755.

7. Agnello L, Vidali M, Padoan A, et al. Machine learning algorithms in sepsis[J]. Clinica

Chimica Acta, 2024, 553: 117738.

8. Islam K R, Prithula J, Kumar J, et al. Machine learning-based early prediction of sepsis using electronic health records: a systematic review[J]. Journal of clinical medicine, 2023, 12(17): 5658.

9. Taylor R A, Pare J R, Venkatesh A K, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach[J]. Academic emergency medicine, 2016, 23(3): 269-278.

10. Zhou L, Shao M, Wang C, et al. An early sepsis prediction model utilizing machine learning and unbalanced data processing in a clinical context[J]. Preventive Medicine Reports, 2024, 45: 102841.

11. Liu S, Ong M L, Mun K K, et al. Early prediction of sepsis via smote upsampling and mutual information based downsampling[C]//2019 Computing in Cardiology (CinC). IEEE, 2019: Page 1-Page 4.

12. Yang H, Spasic I, Keane J A, et al. A text mining approach to the prediction of disease status from clinical discharge summaries[J]. Journal of the American Medical Informatics Association, 2009, 16(4): 596-600.

13. Popejoy L L, Khalilia M A, Popescu M, et al. Quantifying care coordination using natural language processing and domain-specific ontology[J]. Journal of the American Medical Informatics Association, 2015, 22(e1): e93-e103.

14. Yan M Y, Gustad L T, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review[J]. Journal of the American Medical Informatics Association, 2022, 29(3): 559-575.

15.Yang R, Tan T F, Lu W, et al. Large language models in health care: Development, applications, and challenges[J]. Health Care Science, 2023, 2(4): 255-263.

16. Jin M, Yu Q, Zhang C, et al. Health-llm: Personalized retrieval-augmented disease prediction model[J]. arXiv preprint arXiv:2402.00746, 2024, 10.

17. Yu Q, Jin M, Shu D, et al. Health-LLM: Personalized Retrieval-Augmented Disease Prediction System[J]. arXiv preprint arXiv:2402.00746, 2024.

18. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

19. Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey[J]. International journal of computer vision, 2021, 129(6): 1789-1819.

20. Meng H, Lin Z, Yang F, et al. Knowledge distillation in medical data mining: a survey[C]//5th international conference on crowd science and engineering. 2021: 175-182.

21. Jiang S, Dai S, Li Y, et al. Development and validation of a screening tool for sepsis without laboratory results in the emergency department: a machine learning study[J]. EClinicalMedicine, 2025, 80.

22. Prithula J, Islam K R, Kumar J, et al. A novel classical machine learning framework for early sepsis prediction using electronic health record data from ICU patients[J]. Computers in Biology and Medicine, 2025, 184: 109284.

23. Liu Z, Shu W, Li T, et al. Interpretable machine learning for predicting sepsis risk in emergency triage patients[J]. Scientific Reports, 2025, 15(1): 887.

24. Duan Y, Huo J, Chen M, et al. Early prediction of sepsis using double fusion of deep features and handcrafted features[J]. Applied Intelligence, 2023, 53(14): 17903-17919.

25. Sinha S, Kumari K, Kumar A, et al. Multi-Modal Data Fusion for Predictive Analytics in Healthcare: A Hierarchical Attention Framework for Early Sepsis Prediction[C]//2025 6th International Conference for Emerging Technology (INCET). IEEE, 2025: 1-6.

26. Bajpai S, Kaul M. DynEdges-TGN: Dynamic-Edges-Based Temporal Graph Network for Early Sepsis Prediction[J]. SN Computer Science, 2025, 6(3): 285.

27. Zhang W, Wu M, Zhou L, et al. A sepsis diagnosis method based on Chain-of-Thought reasoning using Large Language Models[J]. Biocybernetics and Biomedical Engineering, 2025, 45(2): 269-277.

28. Li Q, Ma H, Song D, et al. Early prediction of sepsis using chatGPT-generated summaries and structured data[J]. Multimedia Tools and Applications, 2024, 83(41): 89521-89543.

29. Shashikumar S P, Mohammadi S, Krishnamoorthy R, et al. Development and prospective implementation of a large language model based system for early sepsis prediction[J]. npj Digital Medicine, 2025, 8(1): 290.

30. Wong A. Knowledge Distillation for Interpretable Clinical Time Series Outcome Prediction[D]. Massachusetts Institute of Technology, 2023.

31. Ibrahim Z M, Bean D, Searle T, et al. A knowledge distillation ensemble framework for predicting short-and long-term hospitalization outcomes from electronic health records data[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 26(1): 423-435.

32. Ji Y, Wang K, Yuan Y, et al. A knowledge-transfer-based approach for combining ordinal regression and medical scoring system in the early prediction of sepsis with electronic health records[J]. Computational Biology and Chemistry, 2024, 113: 108203.