Int. J. Appl. Math. Comput. Sci., 2024, Vol. 34, No. 4, 535–548
DOI: 10.61822/amcs-2024-0036

APPLICATION OF TEXTUAL REPRESENTATION METHODS FOR CLINICAL NUMERICAL DATA IN EARLY SEPSIS DIAGNOSIS

WEIMIN ZHANG [a], LUYAO ZHOU [a], MIN SHAO [b], CUI WANG [b], YU WANG [a,*]

[a] School of Biomedical Engineering
Anhui Medical University
Tanghe Road, 230032, Hefei, China
e-mail: wangyu@ahmu.edu.cn

[b] Department of Critical Care Medicine
First Affiliated Hospital of Anhui Medical University
Jixi Road, 230022, Hefei, China

Sepsis is a severe infectious disease with high incidence and mortality rates worldwide. Early diagnosis of sepsis in newly admitted intensive care unit patients is crucial to reduce mortality and improve patient outcomes. The manual diagnostic methods heavily rely on subjective clinical experience, while traditional machine learning methods require time-consuming feature engineering and the performance is limited by the knowledge acquired from scarce datasets. Therefore, to address the aforementioned issues, this study proposes a novel textual representation method for clinical numerical data, leveraging pre-trained language models from the field of natural language processing for sepsis prediction. Specifically, this study innovatively transforms structured clinical numerical data of patients into unstructured textual descriptions. This transformation reframes sepsis prediction into a text classification task, leveraging the rich prior semantic knowledge embedded in pre-trained language models to enhance prediction performance. The proposed method is validated using real ICU clinical data. When employing RoBERTa-base, it achieved an F1 score of 79.03%, which represents an improvement of five percentage points compared with commonly used machine learning classifiers. The experiments confirmed that the proposed method enhances the performance of early sepsis diagnosis and introduces new insights for clinical diagnosis of sepsis.

Keywords: sepsis diagnosis, text representation, pre-trained language models, machine learning.

1. Introduction

Sepsis is a systemic inflammatory response syndrome caused by pathogenic microorganisms such as bacteria invading the body, which can simultaneously affect multiple organs, leading to organ dysfunction or failure (Singer *et al.*, 2016; Lelubre and Vincent, 2018; Evans *et al.*, 2021). The incidence and mortality rates of sepsis remain high worldwide (Rhee and Klompas, 2020; Rubens *et al.*, 2020), making it the leading cause of infection-related deaths globally and the most fatal condition in Intensive Care Units (ICUs) (Shankar-Hari *et al.*, 2016; Verdonk *et al.*, 2017). Therefore, accurately diagnosing sepsis upon a patient's admission to the ICU is crucial for reducing mortality rates and improving clinical outcomes.

However, the early diagnosis of sepsis remains challenging (Duncan *et al.*, 2021; Gunsolus *et al.*, 2019). Manual diagnosis of sepsis heavily relies on the expertise of clinical professionals, which includes meticulous observation of patients' vital signs and comprehensive analysis of laboratory indicators (Levy *et al.*, 2003; Angus *et al.*, 2001). Nevertheless, subjective factors are inevitably introduced in manual diagnostic methods, leading to variability in the diagnostic accuracy influenced by differences in medical knowledge. This issue is particularly pronounced when dealing with critically ill patients whose conditions are complex and rapidly changing (Watkins *et al.*, 2022). Therefore, researchers have also attempted to develop machine learning models based on objective clinical data to diagnose sepsis (van der Vegt *et al.*, 2023). For example, Du *et al.* (2019) utilized gradient boosting decision trees (GBDTs)

[*] Corresponding author

combined with clinical data, including vital signs and laboratory indicators, to automatically predict sepsis . In another study, a sepsis prediction model was developed using random forests (RFs) based on 20 key predictive variables (Wang *et al.*, 2021). Similarly, Gao *et al.* (2024) utilized RFs to predict mortality among sepsis patients by selecting 38 features from the MIMIC-IV database. However, despite avoiding the subjectivity of manual diagnosis, these existing machine learning methods still have limitations (Fleuren *et al.*, 2020). On the one hand, machine learning methods require time-consuming and labor-intensive feature engineering. On the other hand, and more importantly, machine learning models can only acquire knowledge from training sets, resulting in the performance being limited by the scarcity of training data.

Recently, pre-trained language models (PLMs), such as bidirectional encoder representations from transformers (BERTs) (Devlin *et al.*, 2019), have emerged in the field of natural language processing (NLP). PLMs acquire prior semantic knowledge from large amounts of unlabeled corpora through pre-training tasks like masked language modeling (MLM) and next sentence prediction (NSP). The acquired knowledge is utilized to enhance downstream NLP tasks, which can be considered a form of transfer learning (Li *et al.*, 2020; Coban *et al.*, 2024). Therefore, it is of interest to know whether this prior semantic knowledge can be applied to the early diagnosis of sepsis. Given that the input of a PLM consists of unstructured text sequences, a transformation method is needed to convert structured clinical numerical data into textual descriptions.

This study, for the first time, proposes a novel textual representation method for clinical numerical data, leveraging PLMs from the field of NLP for sepsis prediction. Specifically, the main contributions of this study are as follows:

(i) This study innovatively transforms structured clinical numerical data of patients into unstructured textual descriptions based on a template. All numerical data are transformed into textual descriptions based on thresholds and filled into text templates to generate patients' textual descriptions. It is worth noting that, the transformation method allows for missing values in the clinical numerical data, hence avoiding errors introduced by imputing.

(ii) After obtaining textual descriptions from patients, this study inputs these descriptions into a PLM to predict whether the patient has sepsis based on the representation of the `[CLS]` token. Thus, early diagnosis of sepsis is transformed into a text classification task. Therefore, this method leverages the prior semantic knowledge embedded in the PLM to enhance task performance.

(iii) This study also conducts experiments using real clinical data from ICU to validate the proposed method. The results indicate that employing RoBERTa-base for backbone PLM achieves the highest F1 score at 79.03%. This represents a significant improvement of five percentage points compared with the best-performing machine learning model. Furthermore, t-tests confirm the statistical significance of these results.

The overall structure of this study takes the form of six sections. A brief review of the related work is presented in Section 2. Section 3 deals with the methodology used in this study. The experimental results are presented in Section 4, while the discussion is provided in Section 5. Finally, Section 6 concludes this study with a summary.

2. Related works

Early research on sepsis primarily employed methods involving biomarkers or imaging techniques (cf. Faix, 2013; Stubbs *et al.*, 2013). However, these methods heavily rely on clinical experience, are subjective, and often lack adequate specificity and sensitivity. Consequently, current studies on sepsis increasingly utilize machine learning or deep learning approaches. The widespread application of PLMs in medical diagnostics holds promise for their integration into sepsis diagnosis (Luo *et al.*, 2024; Cichosz, 2023).

2.1. Machine learning. With the advancement of artificial intelligence technology, more researchers are endeavoring to build predictive models for sepsis. These models integrate patients' clinical data and utilize algorithms from machine learning or deep learning for training and optimization to achieve early prediction and risk assessment of sepsis (Deng *et al.*, 2022; Agnello *et al.*, 2023; van der Vegt *et al.*, 2023). For example, van Doorn *et al.* (2021) conducted a single-center retrospective cohort study using the XGBoost model to predict 31-day mortality based on patients' clinical data within two hours. Burdick *et al.* (2020) similarly employed the XGBoost model, gathering clinical data from 270,438 patients to construct a predictive model for sepsis progression within 48 hours. Li *et al.* (2021) utilized case data from the MIMIC-III dataset to establish and compare five machine learning methods: gradient boosting decision trees (GBDTs), logistic regression (LR), K-nearest neighbors (KNNs), random forest (RF), and support vector machine (SVM) for predicting mortality among sepsis patients. García-Gallo *et al.* (2020) also used machine learning algorithms to build a model predicting one-year mortality rate in sepsis patients based on stochastic gradient boosting

(SGB). In addition to machine learning algorithms, researchers have explored deep learning methods for sepsis-related diagnostics. Rafiei *et al*. (2021) designed an intelligent sepsis prediction model using demographic data, vital signs, and laboratory test results, employing long short-term (LSTM) memory networks, convolutional layers, and fully connected layers. Bedoya *et al*. (2020) utilized multi-output Gaussian processes and recursive neural networks to predict sepsis using data from a tertiary hospital's inpatient population. Aşuroğlu and Oğul (2021) introduced the deep SOFA prediction algorithm (DSPA), combining features from convolutional neural networks (CNNs) with random forest (RF) to predict SOFA scores in sepsis patients. While methods based on machine learning and deep learning have made substantial advances compared to traditional approaches, their semantic knowledge is confined to the training set, thereby limiting the performance of the models.

2.2. Pre-trained language models. PLMs consisting of multiple transformer blocks, such as BERT (Devlin *et al*., 2019) and ERNIE (Zhang *et al*., 2019a), can acquire prior semantic knowledge from large-scale unlabelled corpora through pretraining phase and apply this knowledge to downstream tasks. For example, Sangeetha *et al*. (2022) found that COVID-19 can be accurately and efficiently detected and diagnosed using these PLMs, suggesting these low-cost and readily available methods as reliable approaches for COVID-19 diagnosis. Dong *et al*. (2023) proposed a stable and resource-efficient medical diagnostic system based on PLMs. Wang *et al*. (2023) utilized prompt-based PLM fine-tuning methods to enhance Alzheimer's disease (AD) detection. According to the literature review, researchers have not yet explored the use of PLMs for early sepsis diagnosis.

3. Methods

This section first provides a detailed introduction to the textual representation method of clinical numerical data, followed by an explanation of how to utilize PLMs for the diagnosis of sepsis through text classification tasks. The overall workflow of the proposed method in this study is illustrated in Fig.1. First, clinical numerical data of a patient are collected, followed by data cleaning and preprocessing. Then, the data are transformed into text descriptions based on thresholds and a text template to obtain text sequences. Finally, these sequences are input into a PLM to conduct the text classification task based on the representation of the `[CLS]` token, determining whether or not the patient has sepsis.

3.1. Textual representation of clinical numerical data. Given that clinical data are structured and cannot be directly input into a PLM, this study proposes an innovative transformation method to convert structured clinical data into unstructured text descriptions. Initially, intensivists from the First Affiliated Hospital of Anhui Medical University selected 19 clinical features for this study, with descriptions provided in Appendix. The discrete features are listed in the first block of Table 1, while the second block contains continuous features. After data cleaning and preprocessing, a representation of a patient's clinical data is shown in Fig. 2. Subsequently, for continuous features, different descriptions are formulated based on thresholds. For instance, if the threshold for APACHE II score is 20, then the description for APACHE II for the patient in Fig. 2 would be "APACHE II high risk". For discrete features, descriptions are based on their categorical values. For example, if a patient received analgesic treatment, it is directly described as "Received analgesic treatment." The whole rules for transforming numerical clinical data of the 19 features into text descriptions are listed in Table 1. Then, each textual description corresponding to a feature is input into a template to obtain the patient's textual representation. This template, illustrated in Fig. 2, begins with "A patient", followed by sequentially filling in feature descriptions. It is worth noting that this method does not require handling missing values. If data for a certain feature are missing for a patient, the template remains unfilled for that feature. Through these steps, an unstructured text sequence for inputting into PLM is generated.

3.2. Implementing text classification for sepsis diagnosis. After obtaining the unstructured textual description for a patient, this study employs a PLM to perform the text classification task to determine whether the patient has sepsis.[1] Assuming a patient's textual description with length n, we have

$$X = \{x_1, \dots, x_n\}. \tag{1}$$

The input for the PLM are acquired through the following steps:

(i) Following the conventions of BERT and other PLMs, a `[CLS]` token is added to the beginning and a `[SEP]` token to the end of X. The `[CLS]` token is used for subsequent text classification tasks. Thus, the text description sequence is re-represented as

$$X_s = \{\texttt{[CLS]}, x_1, \dots, x_n, \texttt{[SEP]}\}. \tag{2}$$

(ii) The characters cannot be directly input into a PLM. Therefore, the sequences need to be vectorized.

[1] This study utilizes discriminative PLM models such as BERT (Devlin *et al*., 2019), RoBERTa (Liu *et al*., 2019), ERNIE (Zhang *et al*., 2019; Sun *et al*., 2019; 2020; 2021), XLNet (Yang *et al*., 2019), and ELECTRA (Clark *et al*., 2020)

Table 1. Clinical numerical data transformation rules.

Variables	Rules	Textual description
Sex	0	Female
	1	Male
Rehydration test	1	Received rehydration test
Analgesic treatment	1	Received rehydration test
APACHEII	> 20	APACHEII high risk
	≤ 20	APACHEII low risk
SOFA	> 10	SOFA high risk
	≤ 10	SOFA low risk
Heart	> 100	Tachycardia
	< 60	Bradycardia
	else	Normal heart rate
Sbp and Dbp	$Sbp \geq 180$ or $Dbp \geq 110$	Stage 3 hypertension
	$Sbp \geq 160$ or $Dbp \geq 100$	Stage 2 hypertension
	$Sbp \geq 140$ or $Dbp \geq 90$	Stage 1 hypertension
	else	Normal blood pressure
Breath	> 20	Tachypnea
	< 12	Bradypnea
	else	Normal respiration
SpO_2	≥ 95	Normal oxygenation
	else	Hypoxemia
ALB	> 55	Hyperalbuminemia
	< 35	Hypoalbuminemia
	else	Normal albumin level
NA	> 145	Hypernatremia
	< 135	Hyponatremia
	else	Normal sodium level
CL	> 105	Hyperchloremia
	< 95	Hypochloremia
	else	Normal chloride level
LAC	> 2.2	Hyperlactatemia
	< 0.5	Hypolactatemia
	else	Normal lactate level
BUN	> 7.1	Hyperuremia
	< 2.5	Hypouremia
	else	Normal BUN level
PO_2	> 100	Hyperoxemia
	< 75	Hypooxemia
	else	Normal PO_2 level
HB	Female: > 160	Hyperhemoglobinemia
	< 120	Hypohemoglobinemia
	else	Normal hemoglobin level
	Male: > 175	Hyperhemoglobinemia
	< 130	Hypohemoglobinemia
	else	Normal hemoglobin level
CR	Female: > 80	Hypercreatinemia
	< 44	Hypocreatinemia
	else	Normal creatinine level
	Male: > 104	Hypercreatinemia
	< 59	Hypocreatinemia
	else	Normal creatinine level
BMI	> 23.9	High BMI
	< 18.5	Low BMI
	else	Normal BMI

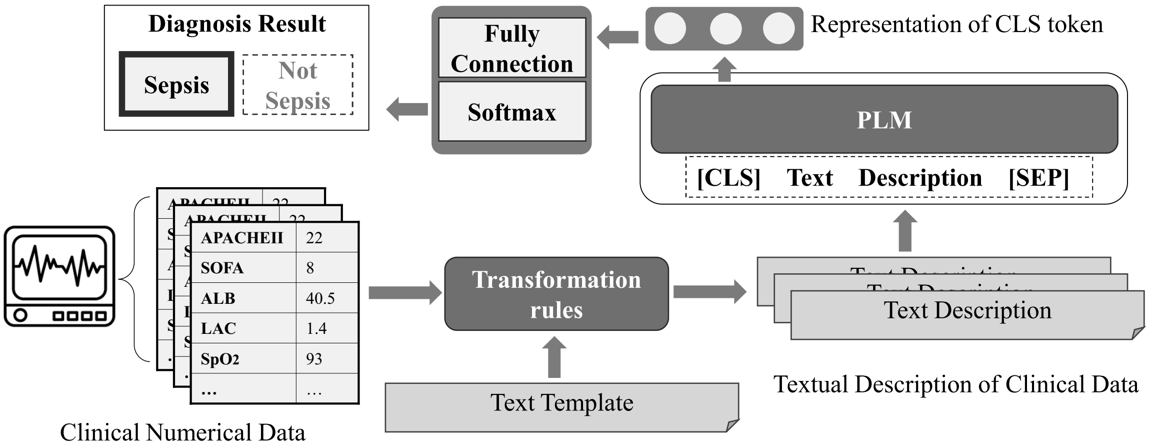

Fig. 1. Flowchart of the method proposed in this study. Clinical numerical data are transformed into textual descriptions, which are then inputted into a PLM to obtain the classification result indicating whether the patient has sepsis.

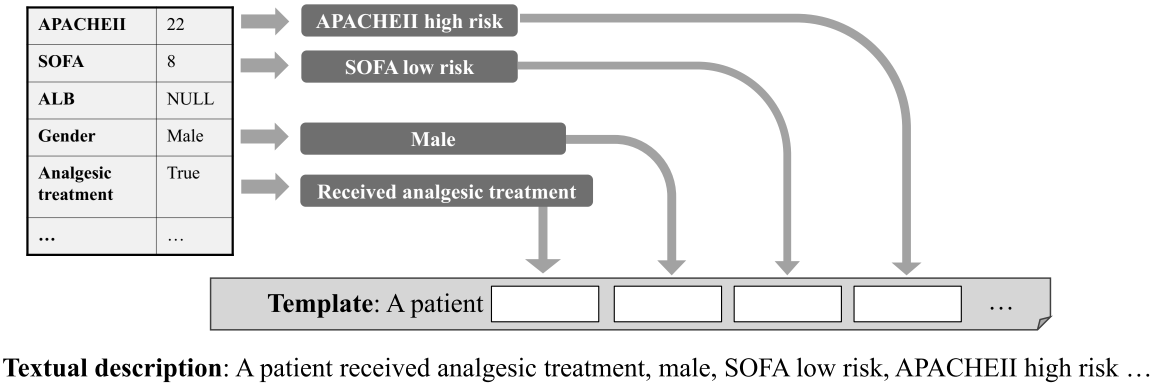
APACHEII 22
SOFA 8
ALB NULL
Gender Male
Analgesic treatment True
… …
APACHEII high risk
SOFA low risk
Male
Received analgesic treatment
Template: A patient …
Textual description: A patient received analgesic treatment, male, SOFA low risk, APACHEII high risk …

Fig. 2. Method for transforming structured clinical numerical data into unstructured textual descriptions. Textual descriptions are obtained for each feature, and then the patient description based on a template is constructed.

Specifically, the X_s is mapped into E_s as the following equation:

$$E_s = \{e(\texttt{[CLS]}), e(x_1), \dots, e(x_n), e(\texttt{[SEP]})\}. \tag{3}$$

The vectorized sequence E_s can be input into a PLM to obtain representations for each token.

For the output generated by pre-trained models, the representation of `[CLS]`, which is regarded as a sentence-level feature, is used to predict the text classification. Specifically, as shown in Fig. 3, the sentence-level category probabilities for token `[CLS]` can be obtained through

$$p_c = \text{softmax}(W_c r_c + b_c), \tag{4}$$

where W_c represents the sentence-level classifier matrix, r_c denotes the column of W_c, and b_c stands for the bias of the classifier. Then, the category can be obtained by

$$y_c = \arg\max(p_c). \tag{5}$$

The loss function for this task is

$$\begin{aligned}\text{loss} = \frac{1}{N}\sum_{i=1}^{N}&(y_c \cdot \log(\sigma(p_c)) \\ &+ (1 - y_c) \cdot \log(1 - \sigma(p_c)),\end{aligned} \tag{6}$$

where N is the number of samples, each sample x_c has a binary label $y_c \in \{0, 1\}$, and the goal of this study is to predict the probability of each sample as $p_c \in [0, 1]$. The sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

transforms any real number x into a probability in the range $[0, 1]$, and log denotes the natural logarithm. The training objective is to minimize the loss function.

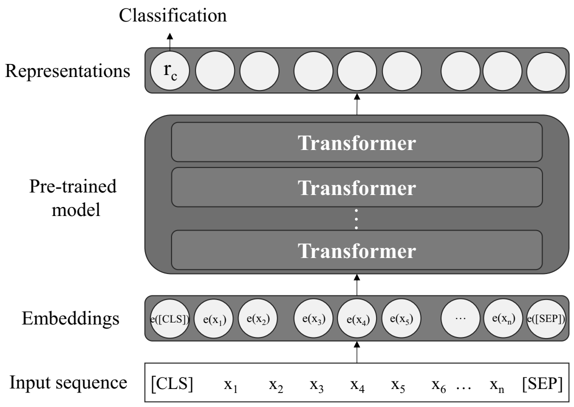

Fig. 3. Text classification using the pre-trained model.

4. Experiments and results

In this section, the dataset used, implementation environment, evaluation metrics, hyperparameters, and experimental results will be detailed.[2]

4.1. Dataset. In this study, the datasets we used were all from a program in collaboration with the First Affiliated Hospital of the Anhui Medical University for Critical Care Sepsis. The data are all from actual ICU clinical environments, and they are authentic and reliable. There were 620 cases in total, which included both sepsis and non-sepsis patients. The baseline analysis of the data is given in Table 2, and we know that $P < 0.05$ indicates statistical significance (Di Leo and Sardanelli, 2020), but we appropriately relaxed the p-value in order to avoid exclusion of key characteristic variables. Finally, we divided the dataset into training and test sets in a ratio of 7:3.

4.2. Implementation environment. For the training and deployment of the PLMs, we utilized the following computational resources: Python 3.7, PaddlePaddle 2.4.0, and PaddleNLP 2.4.2; the hardware configuration included a 4-core CPU, 32GB RAM, V100 GPU with 32GB video memory, and 100 GB HDD.

4.3. Evaluation metrics. For evaluation, this study employed several metrics including accuracy, precision, recall, F1 score, the area under the curve (AUC), and both macro-F1 and micro-F1 for assessing multi-category text classification tasks (Perez-Melo and Kibria, 2020; Cabot and Ross, 2023). Accuracy measures the proportion of correctly classified samples out of the total samples. Precision signifies the ratio of true positive cases correctly classified as positive among all samples classified as positive, while recall indicates the ratio of true positive cases correctly identified as positive among all actual positive cases. AUC represents the area under the ROC curve, quantifying the model's discriminatory capability (Srinivasu *et al.*, 2024). The F1-score is calculated as

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{8}$$

Micro-F1 calculates the overall precision and recall by aggregating the confusion matrices of all classes and subsequently derives the F1-score. It is suitable for handling datasets with class imbalance as it assigns equal weight to each class. By contrast, macro-F1 computes the arithmetic mean of F1-scores across all classes without considering the class distribution, making it suitable for balanced datasets. In this study, genuine ICU clinical data with balanced sample sizes across classes, as depicted in Table 2, were utilized. Therefore, to impartially assess the performance of PLMs in multi-class classification tasks, macro-F1 was chosen as the primary evaluation metric. Macro-F1 effectively measures model classification accuracy across all classes, independent of any class distribution, thereby accurately reflecting the model's overall performance in the task.

4.4. Hyperparameters. The hyperparameters for the model were determined through a trial-and-error approach. To mitigate overfitting and enhance the model's generalization capabilities, we implemented an early stopping strategy during training. This strategy involves monitoring the model's performance on a validation set and halting the training process when the performance stagnates or begins to decline. This approach prevents the model from becoming overly complex and thus overfitting to the training data. The final set of hyperparameters selected is the one that yields the best performance metrics. Table 3 provides a comprehensive list of all the hyperparameters used.

4.5. Experiment results. In this section, we present three aspects of our experimental findings. Firstly, we discuss the performance of machine learning models in a controlled experiment for classification. Subsequently, we provide detailed insights into the text classification predictions of prominent PLMs on the complete dataset, including the loss functions observed during training. Lastly, we analyze the outcomes of conducted statistical tests. Additionally, experiments conducted with varying sample sizes (75%, 50%, and 25%) will be comprehensively detailed in subsequent sections.

4.5.1. Results of machine learning models. This study compares the performance of various common

[2] The source code for verifying this method can be obtained at `https://github.com/bubudai/paper_code.git`.

Table 2. Baseline analysis data.

Variables	Total ($n = 620$)	Non-sepsis ($n = 310$)	Sepsis ($n = 310$)	p-Value
Sex, n(%)				0.740
0	234 (37.7)	115 (37.1)	119 (38.4)	
1	386 (62.3)	195 (62.9)	191 (61.6)	
APACHEII, median[IQR]	19.0 [13.0,24.0]	16.0 [10.0,22.8]	20.4 [16.0,26.0]	0.001
SOFA, median[IQR]	6.0 [4.0,9.0]	5.1 [3.3,8.0]	7.0 [5.0,10.0]	0.001
Heart, median[IQR]	104.0 [82.0,124.0]	93.0 [76.0,112.0]	115.0 [96.0,132.0]	0.001
Sbp, median[IQR]	108.0 [88.0,131.0]	120.0 [100.0,147.0]	98.0 [80.0,120.0]	0.001
Dbp, median[IQR]	61.0 [51.0,76.0]	67.0 [56.0,80.0]	57.0 [46.0,70.0]	0.001
Breath, median[IQR]	20.0 [15.0,26.0]	18.0 [14.0,23.0]	23.0 [17.0,30.0]	0.001
SPO2, median[IQR]	96.0 [92.0,99.0]	97.0 [94.0,99.0]	95.0 [90.0,98.0]	0.001
Rehydration test, n(%)				0.001
0	476 (76.8)	286 (92.3)	190 (61.3)	
1	144 (23.2)	24 (7.7)	120 (38.7)	
Analgesic treatment, n(%)				0.518
0	278 (44.8)	135 (43.5)	143 (46.1)	
1	342 (55.2)	175 (56.5)	167 (53.9)	
alb, median[IQR]	31.4 [27.8,34.5]	32.8 [29.4,36.3]	29.9 [26.0,32.8]	0.001
na, median[IQR]	139.0 [135.9,142.0]	139.3 [136.9,142.0]	138.7 [134.8,142.4]	0.052
cl, median[IQR]	104.0 [99.8,108]	104.4 [100.4,108.1]	103.4 [99.0,108.0]	0.173
lac, median[IQR]	2.2 [1.3,4.5]	1.7 [1.1,3.0]	3.1 [1.8,5.0]	0.001
hb, median[IQR]	106.0 [88.0,127.0]	110.0 [90.0,128.0]	102.0 [87.0,126.0]	0.143
bun, median[IQR]	9.6 [6.2,15.9]	7.3 [5.3,13.5]	12.1 [8.0,17.9]	0.001
cr, median[IQR]	87.3 [61.3,156.2]	73.9 [54.3,112.0]	114.0 [70.6,199.3]	0.001
po2, median[IQR]	100.0 [71.2,135.0]	111.0 [81.0,151.0]	88.0 [65.0,122.6]	0.001
bmi, median[IQR]	22.6 [20.4,24.5]	22.8 [20.8,24.5]	22.5 [20.1,24.5]	0.141

Table 3. Hyperparameters.

Parameters	Value
Maximal length	256
Batch size	64
Learning rate	2e-5
Epoch	20
Optimizer	AdamW

machine learning models in the sepsis prediction task, including support vector machines (SVMs), multi-layer perceptrons (MLPs), k-nearest neighbors (KNNs), XGBoost, decision trees, and random forests (RF). The results, presented in Table 4, show that the accuracy of all models ranges between 65% and 75%, with the random forest model achieving the highest F1 score of 74.42%. ROC curve plots, depicted in Fig. 4, further illustrate the performance of each model across different classification thresholds, demonstrating that common machine learning approaches have some potential for sepsis prediction, but there is room for improvement.

4.5.2. Results of PLMs. This study evaluates the performance of various PLMs in the sepsis prediction task, including ERNIE, XLNet, ELECTRA, BERT, and RoBERTa. The results, presented in Table 5, show that all pre-trained models achieve excellent classification performance, with F1 scores generally above 75%. Notably, the RoBERTa-base model achieves the highest macro-F1 score of 79.03%, significantly outperforming the best result obtained by common machine learning models. Additionally, we analyze the loss functions observed during the training process of different pre-trained models, presented in Fig. 5, to better understand their training dynamics and performance.

4.5.3. Results of *t*-tests. To thoroughly validate the superiority of our proposed method, we meticulously designed ten sets of experiments, each applied to both exceptional pre-trained models and traditional machine learning models. Upon completion of the experiments, we conducted a detailed t-test statistical analysis on the collected data. As depicted in Fig. 6, the t-statistic was remarkably high at 19.10, and the p-value was significantly below the 0.05 (Di Leo and Sardanelli, 2020) threshold (p-value = 2.13e-13). These results clearly indicate a highly significant difference in F1 scores between the RoBERTa-base model and the random forest model, providing us with ample justification to reject the null hypothesis H_0 (which posits no significant difference

Table 4. Classification results of various machine learning models.

Machine models	Accuracy	Precision	Recall	F1 score	AUC
SVM	74.71	77.50	70.46	73.81	75.04
MLP	71.26	73.17	68.18	70.59	73.83
KNN	71.26	72.09	70.46	71.26	72.07
XGBoost	72.41	73.81	70.46	72.09	78.46
Decision Tree	65.52	66.67	63.64	65.12	63.44
Random Forest	74.71	76.19	72.73	74.42	76.04
Logistic Regression	70.12	72.50	65.91	69.05	73.54

Table 5. Classification results of various PLMs.

Pre-trained models	Micro-F1	Macro-F1
ERNIE-1.0-large-zh-cw	75.20	75.15
ERNIE-2.0-base-en	76.36	76.32
ERNIE-3.0-base-zh	75.68	75.66
ERNIE-3.0-nano-zh	76.55	76.55
ERNIE-3.0-mini-zh	77.54	77.52
ERNIE-3.0-micro-zh	76.88	76.87
ERNIE-3.0-medium-zh	76.01	76.01
chinese-XLNet-base	75.68	75.62
chinese-XLNet-mid	76.88	76.88
chinese-XLNet-large	75.20	75.11
chinese-ELECTRA-base	76.68	76.67
chinese-ELECTRA-small	75.27	75.26
BERT-base-chinese	76.68	76.68
BERT-wwm-chinese	75.40	75.38
RoBERTa-large	77.51	77.51
RoBERTa-base	**79.03**	**79.03**

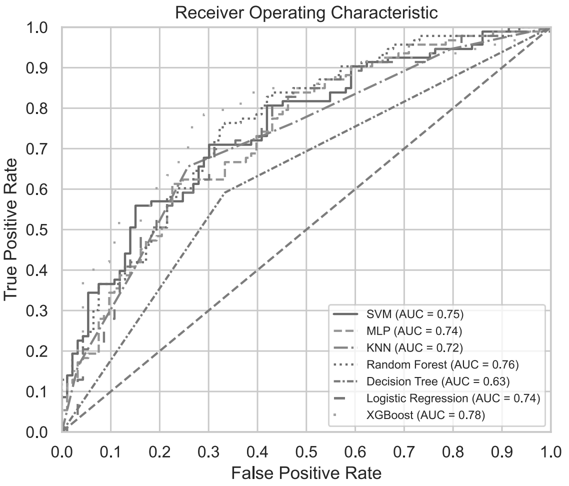

Fig. 4. ROC curves of various machine learning models.

between the two groups). Consequently, we opt to accept the alternative hypothesis H_1 (which posits a significant difference between the two groups), signifying that the text classification approach based on pre-trained models does indeed outperform traditional machine learning methods in the task of sepsis prediction.

4.5.4. Results of experiments with varied sample sizes. In this section, the robustness of the model was further validated by reducing the training set sample size. In the study PLMs were trained using 75%, 50%, and 25% of the training data, respectively, and their performance on the appropriate evaluated test set. As shown in Table 6, even with only 25% of the training set, PLMs maintained an F1 score of 72% or higher, approaching or even surpassing the performance of most common machine learning classifiers on the full training set. These findings clearly demonstrate that PLMs continue to perform well when faced with reduced data volumes, highlighting their superiority and robustness in handling limited data.

5. Discussion

In this section, we will engage in a thorough discussion of the experimental results of our study and highlight some limitations.

5.1. Comparison of different machine learning models. In Section 4.5.1, Table 4 and Fig. 4

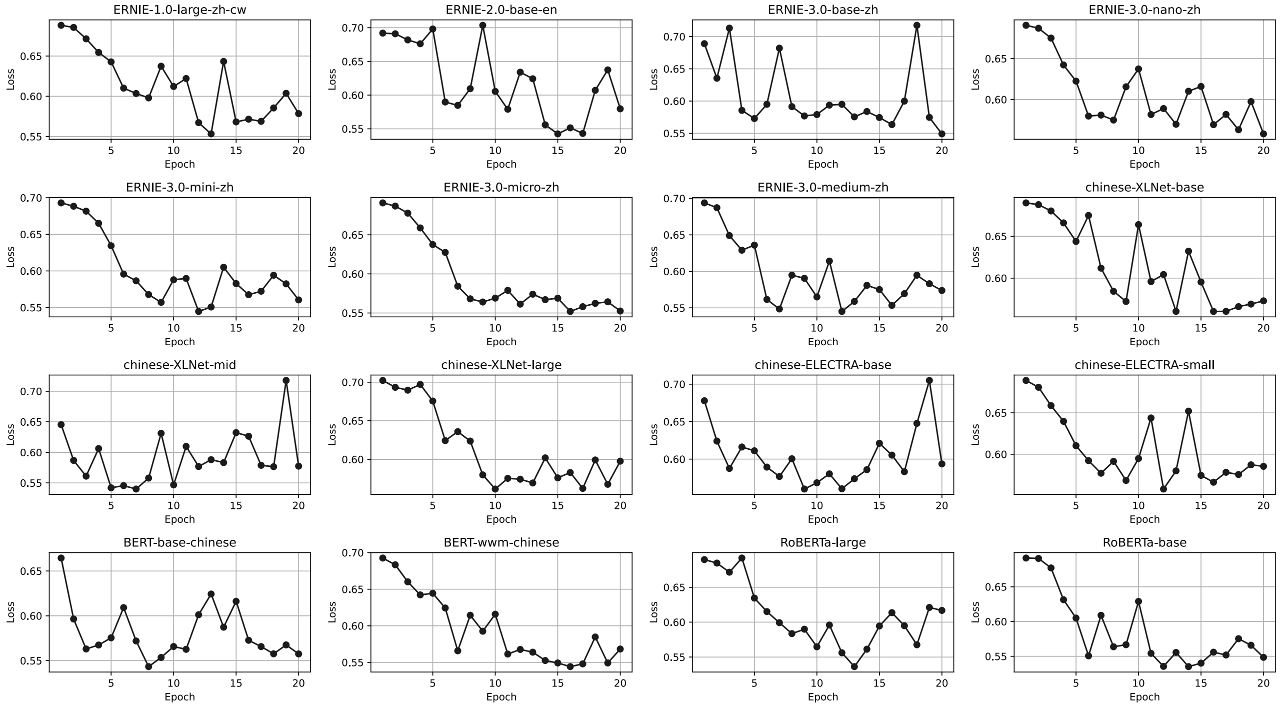

Fig. 5. Comparison of training losses for PLMs in text classification tasks.

various commonly used machine learning models were evaluated for their performance in predicting sepsis. The experimental results indicate that these models exhibit some capability in sepsis prediction, but the overall performance still requires an improvement. The RF model achieved the best performance among all models, with an F1 score of 74.42%, highlighting its effectiveness in feature selection and classification. However, the performance of other models was also relatively close, ranging between 65% and 75%. This suggests that commonly used machine learning models still have limitations in sepsis prediction tasks. These limitations primarily arise from the dependence of machine learning models on training datasets, with their performance being influenced by the dataset size and distribution. In sepsis prediction tasks, data often exhibit complexity and scarcity, posing challenges for model training and prediction. Furthermore, improving model performance depends on effective feature engineering, which requires manual selection and design of features, thereby increasing the difficulty and cost of model development.

5.2. Comparison of different PLMs. In Section 4.5.2, Table 5 and Fig. 5, we evaluated the performance of various PLMs in sepsis classification tasks. The experimental results demonstrate outstanding performance of PLMs in classification. F1 scores generally exceeded 75%, indicating that PLMs effectively leverage their pretraining to enhance the accuracy of sepsis prediction. Among them, the RoBERTa-base model achieved the best performance among all models, with an F1 score of 79.03%, a five-percentage-point improvement over commonly used machine learning models. Its superior performance is primarily attributed to advanced pretraining mechanisms and rich semantic knowledge. RoBERTa enhances BERT's MLM pretraining task with a dynamic masking mechanism, effectively capturing contextual word information. Moreover, RoBERTa benefits from longer training times and larger datasets, further enhancing its semantic understanding and model generalization capabilities.

5.3. Comparison of the performance of t-tests. In Section 4.5.3, Fig. 6, we evaluated the performance of RFs and RoBERTa-base in a sepsis diagnosis task across ten trials. T-tests were conducted on the experimental results, revealing that the RoBERTa-base model achieved significantly higher F1 scores compared with the RF model, with a t-value of 19.10 and a p-value much lower than the significance level of 0.05. This statistical significance indicates that the performance improvement of the RoBERTa-base model is not merely incidental but reflects substantive differences. The heightened statistical significance may be attributed to its deep learning architecture and advantages derived from large-scale pretraining. Compared with traditional machine learning models, RoBERTa-base can capture more intricate

Table 6. Performance of PLMs with different training set sizes.

Sample size	75 Percent training set	50 Percent training set	25 Percent training set
Pre-trained models	Macro-F1	Macro-F1	Macro-F1
ERNIE-1.0-large-zh-cw	74.70	74.03	72.52
ERNIE-2.0-base-en	74.54	72.49	72.69
ERNIE-3.0-base-zh	73.82	73.82	72.33
ERNIE-3.0-nano-zh	77.62	72.92	72.44
ERNIE-3.0-mini-zh	75.47	75.60	72.89
ERNIE-3.0-micro-zh	75.27	74.91	73.27
ERNIE-3.0-medium-zh	76.11	76.25	72.34
chinese-XLNet-base	74.47	76.44	70.99
chinese-XLNet-mid	75.74	77.42	72.87
chinese-XLNet-large	74.69	72.97	72.00
chinese-ELECTRA-base	77.03	74.53	74.41
chinese-ELECTRA-small	74.73	74.71	72.26
BERT-base-chinese	75.32	75.21	72.22
BERT-wwm-chinese	75.27	76.23	72.18
RoBERTa-large	73.66	73.66	72.10
RoBERTa-base	77.06	76.00	73.74

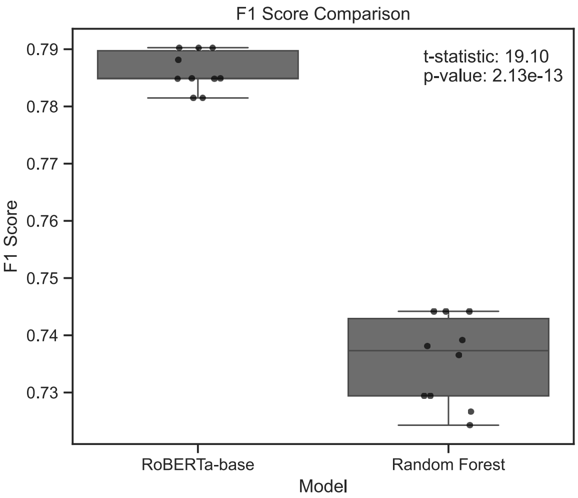

Fig. 6. T-tests of f1 scores between RoBERTa-base and random forest models.

nonlinear relationships and, through extensive pretraining on text data, learn richer feature representations. This deep feature extraction capability enables RoBERTa-base to more accurately identify patterns and biomarkers associated with sepsis when processing clinical text data, demonstrating its potential superiority in clinical applications.

5.4. Comparison of the performance of experiments with varied sample sizes. In Table 6 of Section 4.5.4, we trained PLMs using varying proportions of training data (75%, 50%, and 25%) and evaluated their performance on the corresponding test sets. The results indicate that even with only 25% of the training data, PLMs achieved F1 scores consistently above 72%, approaching or even surpassing the performance of machine learning models trained on the entire training set. This suggests that PLMs maintain strong performance with limited data, demonstrating robustness. The resilience of PLMs to sample size limitations is primarily attributed to their rich semantic knowledge acquired through pretraining on large-scale datasets. Even with reduced training data, PLMs leverage their pretrained semantic knowledge effectively for sepsis prediction. Additionally, having been exposed to diverse textual data during pretraining, PLMs exhibit strong generalization capabilities, adapting well to variations across different datasets.

5.5. Limitations and shortcomings. This study extensively analyzed and compared the performance of different PLMs in early sepsis diagnosis tasks. However, our research still has some limitations. We focused on discriminative PLMs from the BERT series and did not compare them with generative language models like GPT (generative pretrained transformer) and GLM (generative language model). Considering the higher computational demands of large language models (LLMs), we plan to incorporate them into early sepsis diagnosis tasks using CoT (chain of thought) technology in future work.

6. Conclusions

In this study, we explored the potential of leveraging PLMs for sepsis diagnosis. Given that PLMs accept unstructured text sequences as input, our study innovatively proposes a method to transform structured clinical numerical data into unstructured textual representations. This transformation reframes sepsis prediction as a text classification task, utilizing the rich a priori semantic knowledge embedded in PLMs to enhance predictive performance. The study validates this novel approach using real ICU clinical data. Results show that employing RoBERTa-base as the backbone PLM yields the highest F1 score of 79.03%, marking a significant improvement of five percentage points over the best-performing common machine learning model. Moreover, t-tests confirm the statistical significance of these findings. In conclusion, this study advances our understanding of using PLMs for sepsis prediction. Future research will further explore the application of LLMs in this domain. By integrating LLMs, we aim to capitalize on their robust semantic comprehension (Brown *et al.*, 2020) and reasoning capabilities (Wei *et al.*, 2022) to improve the accuracy and efficiency of early sepsis diagnosis.

Acknowledgment

This study was supported by the Research Fund of the Anhui Institute of Translational Medicine (grant no. 2022zhyxC26).

References

Agnello, L., Vidali, M., Padoan, A., Lucis, R., Mancini, A., Guerranti, R., Plebani, M., Ciaccio, M. and Carobene, A. (2023). Machine learning algorithms in sepsis, *Clinica Chimica Acta* **553**: 117738, DOI:10.1016/j.cca.2023.117738.

Angus, D.C., Linde-Zwirble, W.T., Lidicker, J., Clermont, G., Carcillo, J. and Pinsky, M.R. (2001). Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care, *Critical Care Medicine* **29**(7): 1303–1310, DOI:10.1097/00003246-200107000-00002.

Aşuroğlu, T. and Oğul, H. (2021). A deep learning approach for sepsis monitoring via severity score estimation, *Computer Methods and Programs in Biomedicine* **198**: 105816, DOI:10.1016/j.cmpb.2020.105816.

Bedoya, A.D., Futoma, J., Clement, M.E., Corey, K., Brajer, N., Lin, A., Simons, M.G., Gao, M., Nichols, M., Balu, S., Heller, K., Sendak, M. and O'Brien, C. (2020). Machine learning for early detection of sepsis: an internal and temporal validation study, *JAMIA Open* **3**(2): 252–260, DOI:10.1093/jamiaopen/ooaa006.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020). Language models are few-shot learners, *Advances in Neural Information Processing Systems* **33**: 1877–1901, DOI:10.48550/arxiv.2005.14165.

Burdick, H., Pino, E., Gabel-Comeau, D., Gu, C., Roberts, J., Le, S., Slote, J., Saber, N., Pellegrini, E., Green-Saxena, A., Hoffman, J. and Das, R. (2020). Validation of a machine learning algorithm for early severe sepsis prediction: A retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals, *BMC Medical Informatics and Decision Making* **20**(276): 1–10, DOI:10.1186/s12911-020-01284-x.

Cabot, J.H. and Ross, E.G. (2023). Evaluating prediction model performance, *Surgery* **174**(3): 723–726, DOI:10.1016/j.surg.2023.05.023.

Cichosz, P. (2023). Bag of words and embedding text representation methods for medical article classification, *International Journal of Applied Mathematics and Computer Science* **33**(4): 603–621, DOI:10.34768/amcs-2023-0043.

Clark, K., Luong, M.-T., Le, Q.V. and Manning, C.D. (2020). Electra: Pre-training text encoders as discriminators rather than generators, *arXiv:* 2003.10555.

Coban, O., Yağanoğlu, M. and Bozkurt, F. (2024). Domain effect investigation for BERT models fine-tuned on different text categorization tasks, *Arabian Journal for Science and Engineering* **49**(3): 3685–3702, DOI:10.1007/s13369-023-08142-8.

Deng, H.-F., Sun, M.-W., Wang, Y., Zeng, J., Yuan, T., Li, T., Li, D.-H., Chen, W., Zhou, P., Wang, Q. and Jiang, H. (2022). Evaluating machine learning models for sepsis prediction: A systematic review of methodologies, *Iscience* **25**(1): 103651, DOI:10.1016/j.isci.2021.103651.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA,* pp. 4171–4186.

Di Leo, G. and Sardanelli, F. (2020). Statistical significance: p-value, 0.05 threshold, and applications to radiomics-Reasons for a conservative approach, *European Radiology Experimental* **4**(1): 1–8, DOI:10.1186/s41747-020-0145-y.

Dong, B., Wang, Z., Li, Z., Duan, Z., Xu, J., Pan, T., Zhang, R., Liu, N., Li, X., Wang, J., Liu, C., Dong, L., Mao, C., Gao, J. and Wang, J. (2023). Toward a stable and low-resource PLM-based medical diagnostic system via prompt tuning and MoE structure, *Scientific Reports* **13**(1): 12595, DOI:10.21203/rs.3.rs-2313334/v1.

Du, J.A., Sadr, N. and de Chazal, P. (2019). Automated prediction of sepsis onset using gradient boosted decision trees, *2019 Computing in Cardiology (CinC), Baltimore, USA*, p. 1, DOI:10.22489/CinC.2019.423.

Duncan, C.F., Youngstein, T., Kirrane, M.D. and Lonsdale, D.O. (2021). Diagnostic challenges in sepsis, *Current Infectious Disease Reports* **23**(22): 1–14, DOI:10.1007/s11908-021-00765-y.

Evans, L., Rhodes, A., Alhazzani, W., Antonelli, M., Coopersmith, C.M., French, C., Machado, F.R., Mcintyre, L., Ostermann, M., Prescott, H.C., Schorr, C., Simpson, S., Wiersinga, W.J., Alshamsi, F., Angus, D.C., Arabi, Y., Azevedo, L., Beale, R., Beilman, G., Belley-Cote, E., Burry, L., Cecconi, M., Centofanti, J., Coz Y.A., De Waele, J., Dellinger, R.P., Doi, K., Du, B., Estenssoro, E., Ferrer, R., Gomersall, C., Hodgson, C., Hylander M.M., Iwashyna, T., Jacob, S., Kleinpell, R., Klompas, M., Koh, Y., Kumar, A., Kwizera, A., Lobo, S., Masur, H., McGloughlin, S., Mehta, S., Mehta, Y., Mer, M., Nunnally, M., Oczkowski, S., Osborn, T., Papathanassoglou, E., Perner, A., Puskarich, M., Roberts, J., Schweickert, W., Seckel, M., Sevransky, J., Sprung, C.L., Welte, T., Zimmerman, J. and Levy, M. (2021). Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021, *Critical Care Medicine* **49**(11): e1063–e1143.

Faix, J.D. (2013). Biomarkers of sepsis, *Critical Reviews in Clinical Laboratory Sciences* **50**(1): 23–36, DOI:10.3109/10408363.2013.764490.

Fleuren, L.M., Klausch, T.L., Zwager, C.L., Schoonmade, L.J., Guo, T., Roggeveen, L.F., Swart, E.L., Girbes, A.R., Thoral, P., Ercole, A., Hoogendoorn, M. and Elbers, P.W.G. (2020). Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy, *Intensive Care Medicine* **46**: 383–400, DOI:10.1007/s00134-019-05872-y.

Gao, J., Lu, Y., Domingo, I.R., Alaei, K. and Pishgar, M. (2024). Predicting sepsis mortality using machine learning methods, *medRxiv*: 2024.03.14.24304184, DOI:10.1101/2024.03.14.24304184.

García-Gallo, J.E., Fonseca-Ruiz, N., Celi, L. and Duitama-Muñoz, J. (2020). A machine learning-based model for 1-year mortality prediction in patients admitted to an intensive care unit with a diagnosis of sepsis, *Medicina Intensiva* **44**(3): 160–170, DOI:10.1016/j.medin.2018.07.016.

Gunsolus, I.L., Sweeney, T.E., Liesenfeld, O. and Ledeboer, N.A. (2019). Diagnosing and managing sepsis by probing the host response to infection: Advances, opportunities, and challenges, *Journal of Clinical Microbiology* **57**(7): 10–1128, DOI:10.1128/jcm.00425-19.

Lelubre, C. and Vincent, J.-L. (2018). Mechanisms and treatment of organ failure in sepsis, *Nature Reviews Nephrology* **14**(7): 417–427.

Levy, M.M., Fink, M.P., Marshall, J.C., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S.M., Vincent, J.-L. and Ramsay, G. (2003). 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference, *Critical Care Medicine* **31**(4): 1250–1256, DOI:10.1097/01.CCM.0000050454.01978.3B.

Li, K., Shi, Q., Liu, S., Xie, Y. and Liu, J. (2021). Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree, *Medicine* **100**(19): e25813, DOI:10.1097/md.0000000000025813.

Li, X., Zhang, H. and Zhou, X.-H. (2020). Chinese clinical named entity recognition with variant neural structures based on BERT methods, *Journal of Biomedical Informatics* **107**: 103422, DOI:10.1016/j.jbi.2020.103422.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach, *arXiv:* 1907.11692.

Luo, X., Deng, Z., Yang, B. and Luo, M.Y. (2024). Pre-trained language models in medicine: A survey, *Artificial Intelligence in Medicine* **154**: 102904, DOI:10.1016/j.artmed.2024.102904.

Perez-Melo, S. and Kibria, B.G. (2020). On some test statistics for testing the regression coefficients in presence of multicollinearity: A simulation study, *Stats* **3**(1): 40–55, DOI:10.3390/stats3010005.

Rafiei, A., Rezaee, A., Hajati, F., Gheisari, S. and Golzan, M. (2021). SSP: Early prediction of sepsis using fully connected LSTM-CNN model, *Computers in Biology and Medicine* **128**: 104110, DOI:10.1016/j.compbiomed.2020.104110.

Rhee, C. and Klompas, M. (2020). Sepsis trends: increasing incidence and decreasing mortality, or changing denominator?, *Journal of Thoracic Disease* **12**(Suppl 1): S89.

Rubens, M., Saxena, A., Ramamoorthy, V., Das, S., Khera, R., Hong, J., Armaignac, D., Veledar, E., Nasir, K. and Gidel, L. (2020). Increasing sepsis rates in the United States: Results from national inpatient sample, 2005 to 2014, *Journal of Intensive Care Medicine* **35**(9): 858–868.

Sangeetha, S., Kumar, M.S., K, D., Rajadurai, H., Maheshwari, V. and Dalu, G.T. (2022). An empirical analysis of an optimized pretrained deep learning model for COVID-19 diagnosis, *Computational and Mathematical Methods in Medicine* **2022**(1): 9771212, DOI:10.1155/2022/9771212.

Shankar-Hari, M., Phillips, G.S., Levy, M.L., Seymour, C.W., Liu, V.X., Deutschman, C.S., Angus, D.C., Rubenfeld, G.D. and Singer, M. (2016). Developing a new definition and assessing new clinical criteria for septic shock: For the third international consensus definitions for sepsis and septic shock (sepsis-3), *Journal of the American Medical Association* **315**(8): 775–787.

Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.-D. and Coopersmith, C.M. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3), *Journal of the American Medical Association* **315**(8): 801–810.

Srinivasu, P.N., Sirisha, U., Sandeep, K., Praveen, S.P., Maguluri, L.P. and Bikku, T. (2024). An interpretable approach with explainable AI for heart stroke prediction, *Diagnostics* **14**(2): 128, DOI:10.3390/diagnostics14020128.

Stubbs, D.J., Yamamoto, A.K. and Menon, D.K. (2013). Imaging in sepsis-associated encephalopathy—Insights and opportunities, *Nature Reviews Neurology* **9**(10): 551–561, DOI:10.1038/nrneurol.2013.177.

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y. and Lu, Y. (2021). ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation, *arXiv:* 2107.02137.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H. and Wu, H. (2019). ERNIE: Enhanced representation through knowledge integration, *arXiv:* 1904.09223.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H. and Wang, H. (2020). ERNIE 2.0: A continual pre-training framework for language understanding, *Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA*, pp. 8968–8975, DOI:10.1609/aaai.v34i05.6428.

van der Vegt, A.H., Scott, I.A., Dermawan, K., Schnetler, R.J., Kalke, V.R. and Lane, P.J. (2023). Deployment of machine learning algorithms to predict sepsis: systematic review and application of the salient clinical AI implementation framework, *Journal of the American Medical Informatics Association* **30**(7): 1349–1361, DOI:10.1093/jamia/ocad075.

van Doorn, W.P., Stassen, P.M., Borggreve, H.F., Schalkwijk, M.J., Stoffers, J., Bekers, O. and Meex, S.J. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis, *PLoS One* **16**(1): e0245157, DOI:10.1371/journal.pone.0245157.

Verdonk, F., Blet, A. and Mebazaa, A. (2017). The new sepsis definition: Limitations and contribution to research and diagnosis of sepsis, *Current Opinion in Anesthesiology* **30**(2): 200–204.

Wang, D., Li, J., Sun, Y., Ding, X., Zhang, X., Liu, S., Han, B., Wang, H., Duan, X. and Sun, T. (2021). A machine learning model for accurate prediction of sepsis in ICU patients, *Frontiers in Public Health* **9**: 754348, DOI:10.3389/fpubh.2021.754348.

Wang, Y., Deng, J., Wang, T., Zheng, B., Hu, S., Liu, X. and Meng, H. (2023). Exploiting prompt learning with pre-trained language models for Alzheimer's disease detection, *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes, Greece*, pp. 1–5.

Watkins, R.R., Bonomo, R.A. and Rello, J. (2022). Managing sepsis in the era of precision medicine: Challenges and opportunities, *Expert Review of Anti-Infective Therapy* **20**(6): 871–880, DOI:10.1080/14787210.2022.2040359.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems* **35**: 24824–24837, DOI:10.48550/arxiv.2201.11903.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V. (2019). XLNet: Generalized autoregressive pretraining for language understanding, *arXiv:* 1906.08237.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M. and Liu, Q. (2019a). ERNIE: Enhanced language representation with informative entities, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*, pp. 1441–1451.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M. and Liu, Q. (2019b). ERNIE: Enhanced language representation with informative entities. *arXiv:* 1905.07129.

Weimin Zhang received his BS degree from Shandong Youth Politics College in 2023. He is currently an MS degree student of biomedical engineering in Anhui Medical University.

Luyao Zhou received his BS degree from Shandong Youth Politics College in 2022. Her is currently an MS degree student of biomedical engineering in Anhui Medical University.

Min Shao, MD, chief physician, professor, is the director of the Department of Critical Care Medicine, The First Affiliated Hospital of Anhui Medical University. For more than 20 years he has been devoted to basic and clinical research on critically ill multi-organ dysfunction. He was a postdoctoral visiting scholar at Mayo Clinic in the United States from 2014 to 2015, where he was engaged in clinical scientific research on critical care big data.

Cui Wang, attending physician, graduated from Peking Union Medical College in 2020 with a doctorate degree in emergency medicine. She has been engaged in critical care medicine for 10 years, specializing in critical care hemodynamic therapy, critical care ultrasound, and critical care infection therapy, etc. She focuses on the clinical and basic research of sepsis-related hemodynamics and organ function damage.

Yu Wang, PhD, is a school-appointed associate professor and master's supervisor in the Department of Medical Information Engineering, College of Biomedical Engineering, Anhui Medical University, Anhui, China. He graduated from University of Science and Technology of China in 2021, majoring in pattern recognition and intelligent systems. His research interests include the construction of medical knowledge graphs based on unstructured medical literature data, the study of drug recommendation system based on graph neural networks, the design and development of medical big data system, and the methodology of data analytics based on the medical big data platform.

Appendix

Detailed description of abbreviated features

SEX: The gender of the patient.
Rehydration test: Whether the patient received rehydration therapy.
Analgesic treatment: Whether the patient received analgesia.
APACHEII: Acute Physiology and Chronic Health Evaluation II score at the time of ICU admission, used to assess the severity of illness and prognosis.
SOFA: Sequential Organ Failure Assessment score at the time of ICU admission, used to assess the severity of organ dysfunction.
Heart: The patient's heart rate. Unit: beats per minute (bpm).

Sbp and Dbp: The patient's systolic and diastolic blood pressure. Unit: mmHg.

Breath: The patient's respiratory rate. Unit: breaths per minute (bpm).

SpO$_2$: The patient's peripheral oxygen saturation. Unit: %.

ALB: Serum albumin concentration at the time of ICU admission. Unit: g/L.

NA: Serum sodium ion concentration at the time of ICU admission. Unit: mmol/L.

CL: Serum chloride ion concentration at the time of ICU admission. Unit: mmol/L.

LAC: Serum lactate concentration at the time of ICU admission. Unit: mmol/L.

BUN: Serum urea nitrogen concentration at the time of ICU admission. Unit: mmol/L.

PO$_2$: Arterial oxygen partial pressure at the time of ICU admission. Unit: mmHg.

HB: Hemoglobin concentration at the time of ICU admission. Unit: g/L.

CR: Serum creatinine concentration at the time of ICU admission. Unit: μmol/L

BMI: The patient's body mass index, used to assess weight status.

Received: 7 February 2024
Revised: 31 March 2024
Re-revised: 5 July 2024
Accepted: 30 August 2024