



Early prediction of sepsis using double fusion of deep features and handcrafted features

Yongrui Duan¹ · Jiazhen Huo¹ · Mingzhou Chen¹ · Fenggang Hou² · Guoliang Yan³ · Shufang Li⁴ · Haihui Wang³

Accepted: 21 December 2022 / Published online: 17 January 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

Sepsis is a life-threatening medical condition that is characterized by the dysregulated immune system response to infections, having both high morbidity and mortality rates. Early prediction of sepsis is critical to the decrease of mortality. This paper presents a novel early warning model called Double Fusion Sepsis Predictor (DFSP) for sepsis onset. DFSP is a double fusion framework that combines the benefits of early and late fusion strategies. First, a hybrid deep learning model that combines both the convolutional and recurrent neural networks to extract deep features is proposed. Second, deep features and handcrafted features, such as clinical scores, are concatenated to build the joint feature representation (early fusion). Third, several tree-based models based on joint feature representation are developed to generate the risk scores of sepsis onset that are combined with an End-to-End neural network for final sepsis detection (late fusion). To evaluate DFSP, a retrospective study was conducted, which included patients admitted to the ICUs of a hospital in Shanghai China. The results demonstrate that the DFSP outperforms state-of-the-art approaches in early sepsis prediction.

Keywords Sepsis prediction · Early fusion · Late fusion · Deep learning · Tree-based model

1 Introduction

The immune system of the body plays an important role in preventing and limiting infection caused by bacteria, viruses, parasites, or fungi. Sepsis is a severe immunological reaction to infection that can cause tissue damage and organ dysfunction. This reaction can proceed to septic shock, which includes organ failure and very low blood pressure [1]. The Systemic Inflammatory Response Syndrome (SIRS) was proposed in 1991 as the first definition of sepsis [2]; while in 2001, the International Sepsis Definitions Conference revised the definition of sepsis (Sepsis-2) that has facilitated the physician to diagnose sepsis at the bedside [3]. The Third International Consensus Definitions for Sepsis (Sepsis-3) was published in 2016, with a new definition of sepsis and septic shock [4]. In Sepsis-3, Sequential Organ Failure Assessment (SOFA) score and

quick SOFA (qSOFA) score are recommended to represent a patient's organ dysfunction.

Sepsis is still a significant global health problem today. Looking at the United States scenario, clinical indicators of sepsis are present in 6% of hospitalized patients, and sepsis is responsible for nearly 35% of all hospital deaths [5]. From an economical point of view, sepsis is the most expensive condition treated in the United States, accounting for \$38.2 billion in hospital costs in 2017 [6]. Given that the treatment process for sepsis is highly time-sensitive, early detection of sepsis is critical for improving septic patient survival [7, 8]. Based on international guidelines for sepsis management [9], the early administration of antibiotics and intravenous fluids is recommended. With a one-hour delay in the administration of intravenous antibiotics, sepsis mortality can rise between 4%–8% [10]. Furthermore, implementing a standardized sepsis treatment process with practical operational constraints takes some time [11], and as a result, it is critical to predict the onset of sepsis early to schedule and carry out a sepsis treatment plan.

The challenge in an earlier prediction of sepsis is distinguishing sepsis from other disease states with similar clinical signs, symptoms, and molecular manifestations, such as inflammation [12]. Due to the systemic character of sepsis, some sepsis biomarkers, have been proposed to

✉ Mingzhou Chen
2010361@tongji.edu.cn

be utilized in sepsis detection [13]. However, this kind of method is not commonly accepted for sepsis detection due to a lack of specificity [13].

Over the last decade, a rising number of available electronic medical records (EMRs) have created several sepsis prediction algorithms based on machine learning. Machine learning is a branch of artificial intelligence, and can overcome the limitations of standard clinical statistical approaches in interpreting high-dimensional, nonlinear, and longitudinal EMR data. The InSight algorithm, one of the earliest developed machine learning models for sepsis early prediction, was introduced by Calvert et al. [14]. Subsequently, Desautels et al. [15] and Mao et al. [16] proposed a prediction model based on the InSight algorithm. Goh, et al. [11] proposed the SERA algorithm, which combines both structured data and unstructured clinical notes to predict sepsis. More recently, deep learning models and tree-based models have been the most used solutions to sepsis prediction problem. A deep learning model is concerned with the representation of learning with neural networks with several layers, which can extract the knowledge of EMRs via hierarchical architecture. For example, Kok et al. [17], Zhang, et al. [18] and Wang and Yao [19] explored deep learning models to early predict sepsis onset. Scherpf et al. [20], Kam and Kim [21], and Fagerstrom, et al. [22] developed the prediction model based on the recurrent neural network (RNN) which presents good results in handling temporal information. Shashikumar, et al. [23] stacked the deep learning model and the modified Weibull Cox Proportional Hazards model to maintain the predictive performance of the sepsis predictor, while improving the interpretability. Unlike a deep learning model, a tree-based model predicts sepsis earlier, by relying on clinical experts to provide handcrafted features that contain valuable information. So far, the tree-based models, such as random forest [24], gradient boosting decision tree (GBDT) [25] and XGBoost [26], have been developed using handcrafted features to effectively predict early sepsis.

As already mentioned, various machine learning models have been used as tools for the early prediction of sepsis, but this task is still challenging. Existing models usually predict sepsis using a feature representation based on either temporal information derived from EMRs or clinical prior knowledge of sepsis. However, temporal information and clinical knowledge render different functions: the former can reflect an immediate trend of output signals at the current time, while the latter is used to depict the physiological state of the patient. That is, the useful information contained in the feature representation derived from the previous sepsis prediction models is insufficient, limiting the predictive performance of these models.

Fusion technology is an effective process of fusing complementary information that can be used for the

overall comprehension of a phenomenon. There are two typical fusion strategies: early fusion and late fusion. Early fusion, as the name implies, is performed at the feature level, i.e., several features are concatenated into a single representation; whereas late fusion is performed at the scores level, i.e., multiple classification results are combined [27]. Fusion technology has recently been used in the healthcare field, due to its higher predictability. Sun, et al. [28] proposed a multimodal deep neural network for breast cancer prognosis prediction, which used late fusion to integrate multi-dimensional data. To handle the time series classification problem such as early diagnosis, Lv, et al. [29] proposed a dynamic late fusion strategy that fuses the predicted results of multiple-based classifiers, having also developed an adaptive learning method to output final prediction results based on dynamic late fusion. Hagerty, et al. [30] employed the fusion of handcrafted features and deep features to achieve a higher accuracy for melanoma diagnosis. Zhang and Chen [31] proposed a view fusion module for human pose estimation, which combines decision-level information from different stages so a more comprehensive estimation could be generated. Ilhan, et al. [32] developed a computer-aided diagnosis system for early diagnosis of COVID-19, which fuses deep features from seven convolutional neural networks (CNN) architectures, feeding them to multiple classifiers using a late fusion strategy. Zuo, et al. [33] proposed a deep multi-fusion framework with classifier-based feature synthesis, which can automatically fuse multi-modal medical pictures, to aid in precision diagnosis and surgery planning in clinical practice. To predict COVID-19 patient health for early monitoring and effective treatment, Gumaei, et al. [34] proposed a fusion technique that combines three well-calibrated ensemble classifiers. Wang, et al. [35] proposed a neural network framework based on multi-view fusion to automatically segment the gross tumor volume in brain glioma. As stated in the previous studies, fusion techniques are efficient at extracting meaningful patterns from multiple complementary information sources. This means that fusion techniques have great potential to combine temporal information and clinical knowledge to further improve the performance of sepsis prediction. However, there is a scarcity of studies on sepsis prediction by using fusion technology. The goal of this paper is to bridge the gap and establish an effective fusion technology to improve the predictive performance and robustness of the sepsis prediction model.

To achieve this goal, it is proposed a novel early sepsis prediction model called Double Fusion Sepsis Predictor (DFSP), which is a double fusion framework that combines the benefits of early and late fusion. First, the temporal information of the EMRs is efficiently extracted, by using a hybrid deep learning model. Second, the handcrafted

features with clinical knowledge are collected, and third, the early fusion strategy is used to combine deep features and handcrafted features for generating joint feature representation. Afterwards, multiple tree-based models based on the joint feature representations are constructed. Finally, a late fusion strategy is used to integrate these tree-based models to output the final prediction. The main contributions of this paper are as follows:

- (1) A novel double fusion framework for sepsis prediction is proposed. To thoroughly and accurately assess the patient's condition, DFSP uses early fusion to establish an informative feature representation that comprises temporal information and clinical knowledge. Regarding late fusion, it is used to eliminate the randomness of the informative feature representation to improve the robustness of the DFSP. To the best of the authors' knowledge, this is the first study that combines early and late fusions to explore the sepsis prediction problem.
- (2) To extract rich and important temporal information, a hybrid deep learning model that combines CNN and RNN is proposed. Unlike standard sepsis prediction models, which usually utilize a single deep learning module to capture temporal information, the hybrid deep learning module first uses CNN and later the RNN. The first is used to identify the features that effectively describe the state of sepsis patients, and the second, the RNN module is utilized to capture key long-term temporal dependencies in EMRs data based on the CNN modality.
- (3) The DFSP model was applied to a retrospective study of infection patients admitted to the ICU of a hospital in Shanghai China. The effectiveness of DFSP is assessed by comparing it with state-of-the-art methods and traditional sepsis detection tools. The experiment results show that DFSP has a significantly higher area under curve (AUC) score than the existing sepsis models.

The remainder of this paper is organized as follows. In Section 2, the dataset and DFSP are provided. In Section 3, the experimental results are listed and analyzed. Finally, Section 4 concludes this study.

2 Materials and methods

2.1 Data

2.1.1 Data description

The dataset used in this work is from a hospital in Shanghai. Between 2016 and 2021, the records of 282 ICU patients were collected from the Infection department of

the Shanghai hospital. All the considered patients were admitted to the ICU with a diagnosis of infection, and 145 of them were diagnosed with lung infection. The features in the dataset included vital signs, laboratory test results and demographics. The details of the used features are given as follows: (1) Vital signs: heart rate, respiratory rate, the state of ventilator usage, systolic blood pressure (SBP), diastolic blood pressure (DBP), mean arterial pressure (MAP), temperature, pulse oximetry (O2Sat), fluid intake, and fluid output; (2) Laboratory test results: partial thromboplastin time, the time of penicillin usage, aspartate transaminase, calcium, bicarbonate, creatinine, C-reactive protein, interleukin-6 (IL-6), potassium, sodium, partial pressure of carbon dioxide from arterial blood (PaCO₂), fraction of inspired oxygen (FiO₂), blood pH, platelets counts, procalcitonin, white cell count, and total bilirubin; (3) Demographics: age, gender, the ICU length of stay (ICULOS). In this dataset, the diagnosis of sepsis is based on Sepsis-2, which is diagnosed as the presence of at least two SIRS criteria and a confirmed or suspected infection. SIRS criteria are defined as:

- Temperature >38°C or <36°C
- Heart rate >90/min
- Respiratory rate >20/min or PaCO₂ <32mmHg
- White cell count >12 cells/uL or <4 cells/uL

2.1.2 Data preprocessing

Before each instance is fed into the DFSP, irregularity of time series and missing values were addressed. These issues are addressed in the preprocessing step.

In the first step of data preprocessing, the time bucket technique is employed to address the irregularity of time series, which can aggregate data by time interval. The raw data is grouped into a series of consecutive 1 h buckets. After that, the measurement values are averaged within each bucket. As a result, each time series is at 1 h intervals.

Then, the missing values in the data are imputed. In the data imputation phase, the time series with missing values are imputed one by one. Linear interpolation is used first, which is a useful method for curve fitting using linear polynomials. The forward-fill and backward-fill are then used to fill the last or first available data value. Following that step, there are still some missing values because not all clinical features are collected for each patient and interpolation cannot be used for missing features. Thus, the remained missing values are subsequently set to zero. The steps of data preprocessing are summarized in Table 1.

2.2 Design and implementation of DFSP

DFSP is divided into two phases: early fusion and late fusion. In the early fusion phase, the multivariate time series

Table 1 The summary of data preprocessing steps

| Issue | Method |
|---|---|
| Irregularity of time series: The different vital signs and laboratory test results are recorded at irregular intervals and varying frequencies. | Time bucket technique |
| Missing values: Due to disturbances in the medical systems, there are missing values in the dataset. | Linear interpolation, forward-fill, and backward-fill |

are fed into several deep learning models, each being a hybrid of a CNN and an RNN. Then, the deep features are extracted for each deep learning model, and finally, the joint feature representation of deep features and handcrafted features are built. In the following phase, several GBDTs are constructed based on the joint feature representation, and the risk of sepsis onset is calculated with a late fusion strategy integrating the output of tree-based models. The main framework of DFSP is shown in Fig. 1.

2.2.1 Convolutional neural network

CNN is a popular deep feed-forward neural network that is good at processing high-dimensional data. The main block of CNN is the convolutional layer, which is used to subsample and extract features. The convolutional layer computes a dot product between the input data and the filter matrix, and the result of the dot product is loaded into an output matrix. The activation function is then applied to the value in the output matrix. In this paper, the rectified linear units (ReLU) activation function is used and is calculated by (1):

$$ReLU(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{1}$$

2.2.2 Gated recurrent unit

Gated recurrent unit (GRU) is a powerful and simplified version of LSTM (Long short-term memory), which can improve network performance with less training time [36]. The structure of a GRU cell is similar to the structure of an LSTM. A GRU merges the input and forget gates of an LSTM into the update gate, and a GRU cell combines the cell state and hidden state into one state. The hidden state is described by (2):

$$h_t = (1 - z_t) * h_{t-1} + z_t * \bar{h}_t \tag{2}$$

where h_{t-1} and \bar{h}_t are the previous and current candidate memory contents, respectively. z_t is the update gate that is calculated by using (3), and that decides how much of

the previous memory contents should be passed along to the future timestep, and how much of the current candidate memory contents to be added:

$$z_t = \sigma (W_z [h_{t-1}, x_t]) \tag{3}$$

where σ is the sigmoid function and W_z is a weight vector which can be learned during the training. The calculations of the reset gate r_t and candidate memory contents \bar{h}_t are described by the following equations:

$$r_t = \sigma (W_r [h_{t-1}, x_t]) \tag{4}$$

$$\bar{h}_t = \tanh (W_c [r_t * h_{t-1}, x_t]) \tag{5}$$

where W_r and W_c are weight vectors.

2.2.3 Gradient boosting decision tree

GBDT is an ensemble machine learning algorithm for both classification and regression problems, which can generate a prediction model by combining a series of classification and regression trees (CARTs). In the GDBT algorithm, the CARTs are constructed iteratively, and a new CART is trained from the prediction error of the previous iteration. Finally, the output is calculated by accumulating the predictive results of all CARTs.

Suppose that the training set is $S = \{(x_1, y_1) \dots (x_n, y_n)\}$, $f(x)$ is a linear combination of CARTs, and the $L(y, f(x))$ is the loss function. The maximum number of CARTs is considered as M . Equation (6) represents the GBDT:

$$f(x) = \sum_{m=1}^M D(x; \theta_m) \tag{6}$$

where $D(x; \theta_m)$ is the m th CART, and the θ_m is the optimal parameter of m th CART. The negative gradient of the loss function $f_m(x)$ is calculated by (7):

$$R_{mi} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=f_{m-1}(x)}, \quad \forall i \in n \tag{7}$$

The m th CART is trained by using $\{(x_1, R_{m1}) \dots (x_n, R_{mn})\}$, thus θ_m is calculated by (8):

$$\hat{\theta}_m = \arg \min_{\theta_m} \sum_{i=1}^n L(y_i, f_{m-1}(x) + D(x; \theta_m)) \tag{8}$$

Then, the function $f_m(x)$ is updated by adding a new CART and it is described by (9):

$$f_m(x) = f_{m-1}(x) + D(x; \theta_m) \tag{9}$$

Finally, the GBDT model is obtained and the output of GBDT, which is calculated by (6).

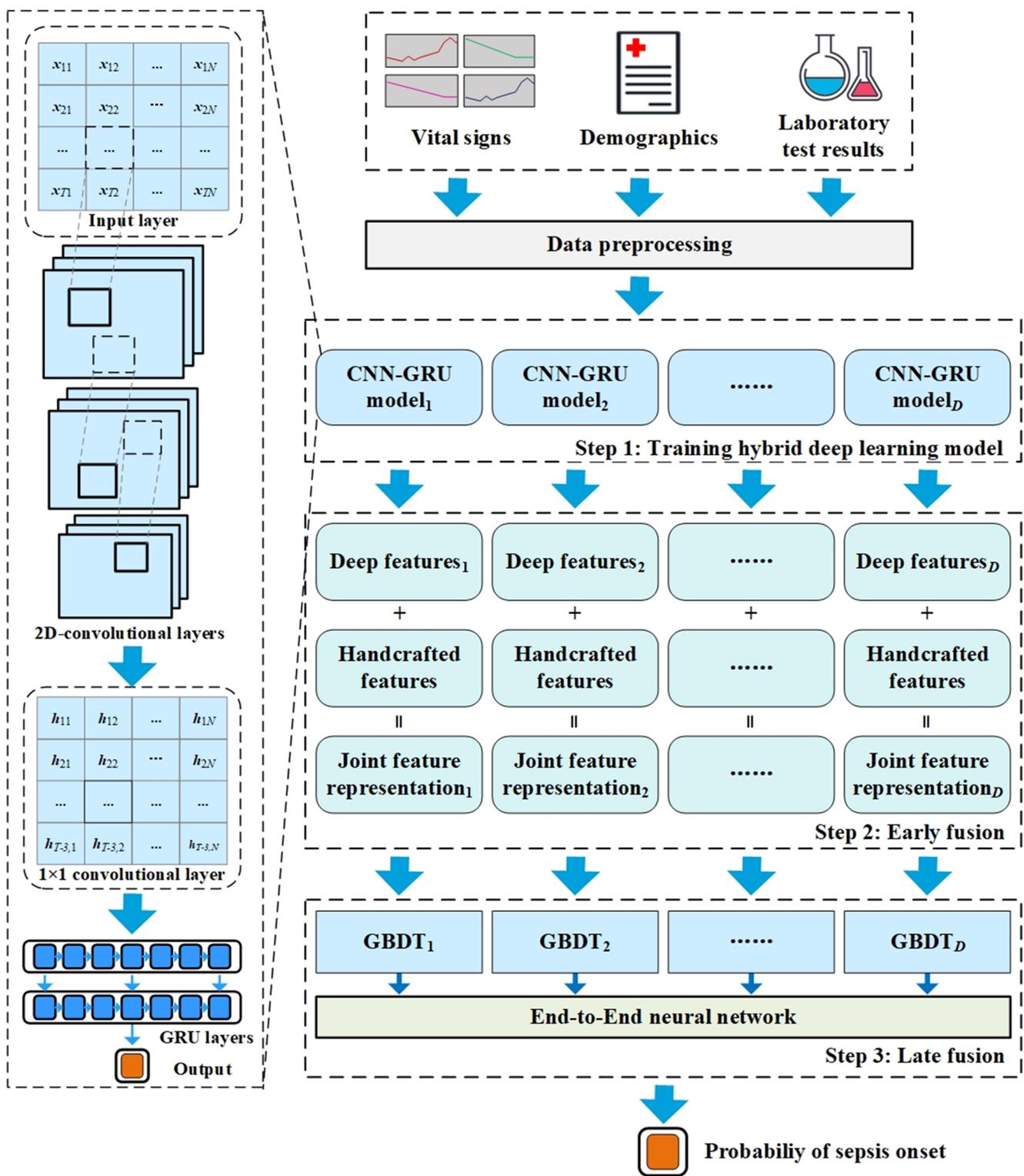


Fig. 1 The main framework of DFSP

2.2.4 Hybrid deep learning model

In this subsection, the structure of the hybrid deep learning model is presented. The hybrid model is composed of two

components: a CNN module for feature extraction, and two GRU layers for prediction based on extracted features. In the first component, three 2D-convolutional layers are employed to extract local features from the time series.

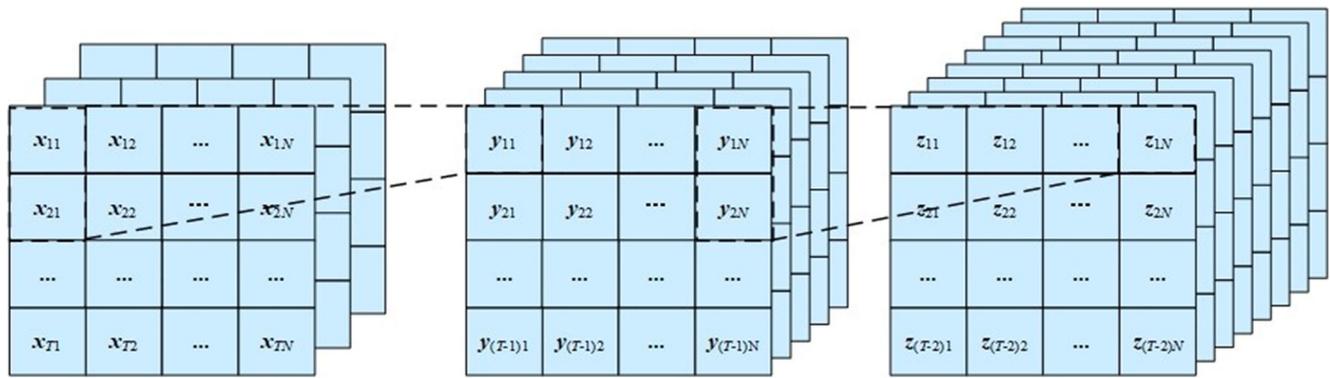


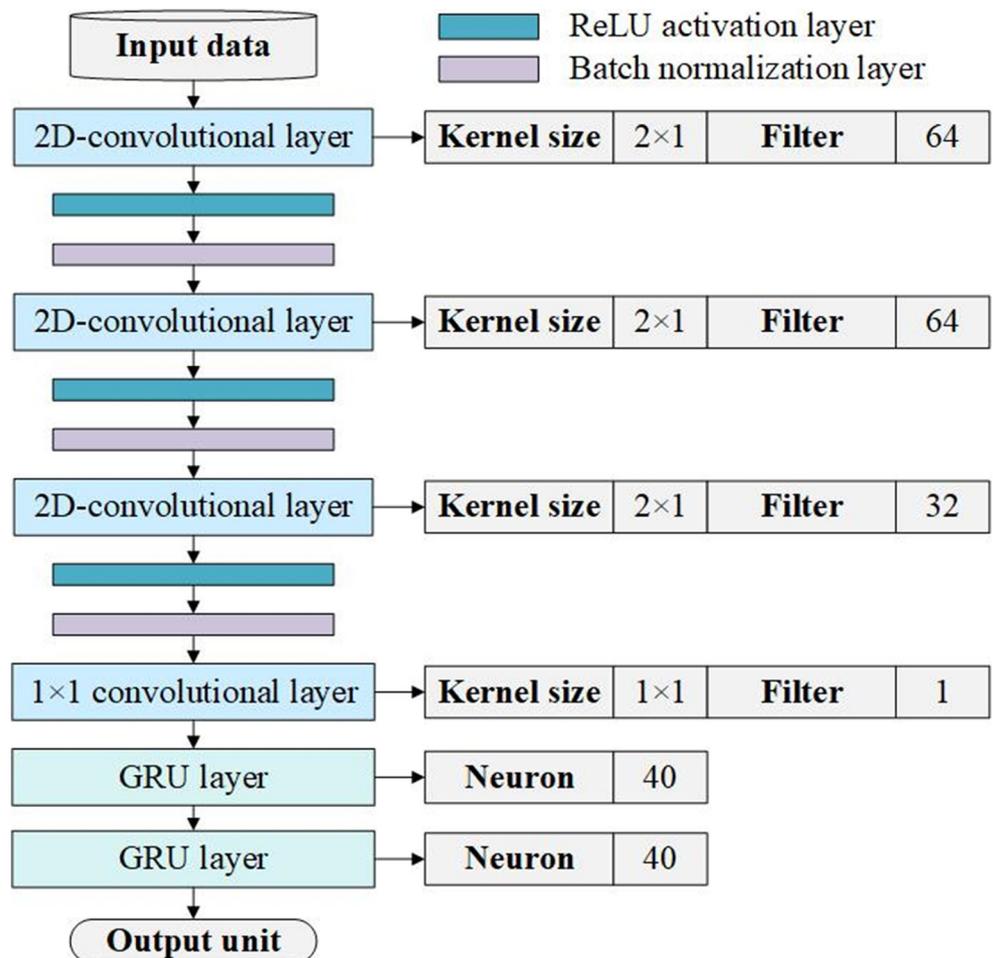
Fig. 2 The 2D-convolutional layers of the hybrid deep learning model

The first 2D-convolutional layer receives the input data and converts it to a feature map. The second layer intensifies the salient feature on the feature maps generated by the first layer, being followed by a third layer that repeats the second layer’s operation on the feature maps, that are generated by the second layer. In the 2D-convolutional layers, a specific kernel size is employed to learn the dynamics of each feature. As shown in Fig. 2, the kernel size is set to 2×1

and thus the convolutional layers are forced to learn the important information from individual time series features.

The architecture of a CNN module also includes batch normalization and rectified linear units (ReLU) activation function. The batch normalization is used to speed up the training process, with the standardization of the feature map. The ReLU activation function is used to compute the output of the normalized feature map. It is important to note that the

Fig. 3 The architecture of the proposed hybrid deep learning model



GRU layer can only read one- or two-dimensional data, but the output of a two-dimensional convolutional layer is three-dimensional data, which is made up of multiple feature maps. In this paper, the extracted features are compressed by using a 1×1 convolutional layer before being fed to the GRU layers.

In the second component, the compacted feature map is fed into the GRU layers, and prediction is performed only based on it. The GRU layers are used to capture the temporal information from the CNN module. Finally, the network computes the probability of sepsis onset by using a single output unit:

$$\text{probability}(x) = \text{sigmoid}(wx + b) \quad (10)$$

where w and b are trained parameters, and x is the output of the last GRU layer. Figure 3 presents the architecture of the proposed hybrid deep learning model.

As shown in Fig. 3, three 2D-convolutional layers have 64, 32, and 32 filters with kernel size 2×1 , respectively. The 1×1 convolutional layer has 1 filter with kernel size 1×1 . The neuron size of each GRU layer in the hybrid deep learning model was set to 40. L1-L2 regularization was employed to reduce overfitting and to enhance model generalizability, with the L2 regularization value set to $1e-5$ and the L1 regularization parameter set to $1e-6$. Due to the class imbalance issue in the dataset, oversampling technology, and noise injection were used in the training procedure of the hybrid deep learning model. A training instance for the deep learning model is made up of the multiple medical time series as input features and the sepsis event indicator as a classification label. The sepsis event indicator is a binary variable. When sepsis occurs within the prediction windows, the sepsis event indicator is set to 1; otherwise, it is set to 0. To optimize the network weights of the hybrid deep learning model, the Adam optimizer [37] is used with a learning rate at $1e-3$. The hybrid deep learning model was trained over a total of 50 epochs with the mini-batch size fixed at 512 in each epoch. The hybrid deep learning model employs the binary cross entropy loss as a loss function. A grid search is used to fine-tune the hyper-parameters of the hybrid deep learning model.

2.2.5 Handcrafted feature

In the subsection, handcrafted features used in DFSP are given, which are made up of three parts: Original feature, statistical feature, and clinical score.

- The value of each feature collected from its most recent record makes up the original features.
- Statistical features are broadly adopted in the machine learning field to improve the predictive performance of models. Statistical features can be used to capture

temporal characteristics while avoiding the data missing problem. Many statistical features are computed by compressing information over time series intervals, e.g., “the standard deviation of the heart rate between 10 and 16 hours.” In this paper, selected statistical features were chosen, such as measurement frequency, mean value, standard deviation, minimum value, and maximum value, and calculated in advance at 6-, 12-, and 24-hour intervals.

- Domain-specific features, which are typically developed by domain specialists and encompass a richness of domain knowledge, can assist prediction models in obtaining superior predictive performance. The clinical score is a type of useful domain-specific feature in the field of disease prediction. In this paper, four sepsis scores were used, including the SOFA score, the qSOFA score, the SIRS criterion, the modified early warning score, and two clinical indexes including the shock index and oxygenation index.

In total, a higher number of 411 handcrafted features were developed.

2.2.6 Double fusion framework

In this subsection, the proposed early and late fusion strategies are given. Early fusion combines multiple relevant features into a single feature vector that contains more information than the initial input feature vectors [38]. In this paper, early fusion is used to combine deep features and handcrafted features for building an informative joint

Table 2 The distribution of joint feature representation

| Feature name | The number of features |
|------------------------|------------------------|
| Original feature | 30 |
| Vital signal | 10 |
| Laboratory test result | 17 |
| Demographic | 3 |
| Statistical feature | 375 |
| Measurement frequency | 51 |
| Mean value | 81 |
| Standard deviation | 81 |
| Maximum value | 81 |
| Minimum value | 81 |
| Clinical score | 6 |
| Sepsis score | 4 |
| Clinical index | 2 |
| Deep feature | 40 |
| Total | 451 |

feature representation. The proposed hybrid deep learning model is trained first, and then its network parameters are frozen. Subsequently, the deep features are extracted from the output sequence of the last GRU layer. The deep features and handcrafted features are fused into a joint feature representation by concatenating them. Thus, the length of joint feature representation is the sum of the number of handcrafted features and the neuro size of the last GRU layer. The distribution of joint feature representation is listed in Table 2. The joint feature representation is fed into GBDT for prediction. Figure 4 illustrates the detailed process of early fusion.

The training process of the deep learning model and that of the GBDT are different, with the former using gradient-based training algorithms and the latter building tree models iteratively. As a result, the training process of the two models needs to be carried out separately. The hybrid deep learning model is trained first to generate deep features, and then the GBDT is trained based on the joint feature representation. Thus, several constructed joint feature representations may not well suit for GBDT resulting in a bad performance, as the deep features cannot be updated with the training loss of GBDT. An intuitive way to solve this problem is to construct various joint feature representations to build various GBDTs and then combine the better GBDTs for final prediction. Due to the randomness inherent in the deep learning method, we can train multiple hybrid deep learning models to obtain various deep features. Then, to get more accurate prediction results, we integrate the advantages of multiple GBDTs in a process called late fusion. Late fusion is an ensemble strategy for

producing more precise and reliable decisions by combining the decisions of multiple classifiers [39]. For traditional late fusion strategies, multiple classification scores are generated, and the scores are merged using a human-created rule. The designs of these human-created rules, such as the sum weight rule, need a significant amount of trial and error. To overcome this disadvantage, an End-to-End neural network was implemented, using classification scores of tree-based models as inputs for training and its output as the final judgment. The End-to-End neural network has only one hidden layer, which is a fully connected layer with 6 neurons. The training process of DFSP is as follows:

- Step 1: Let D denote the number of hybrid deep learning models.
- Step 2: Train the D hybrid deep learning models on patient longitudinal data. All training data will be used to train each hybrid deep learning model.
- Step 3: Freeze the hybrid deep learning model's network weight and feed training data to the hybrid deep learning model to obtain deep features.
- Step 4: Concatenate D pairs of the deep features and handcrafted features to create the D joint feature representations.
- Step 5: Train the D GBDTs with joint feature representations and compute classification scores of D GBDTs in each training data for End-to-End neural network training.

The framework of our double fusion is shown in Fig. 5.

2.3 Evaluation metrics

The major performance indicator is the AUC score, which assesses a model's ability to distinguish between sepsis and non-sepsis patients. Furthermore, measures such as accuracy (Acc), sensitivity (Sens), specificity (Spec) and likelihood ratios (LHR+, LHR-) were also used to analyze the model further. The ability of a model to correctly identify patients is measured by its accuracy. Sensitivity assesses a model's ability to accurately identify patients with sepsis, whereas specificity assesses a model's ability to correctly identify patients who do not have sepsis. The likelihood ratio is the ratio of the likelihood of a specific test result in people with the disease to the likelihood in people without the disease [40]. Prior to this study, numerous studies on sepsis prediction employed alternate metrics such as positive predictive value and negative predictive value to present the probability of diagnosis. However, the prevalence of sepsis in the population substantially influences both positive predictive value and negative predictive value [41]. To alleviate the difficulties associated with interpreting predictive values, Fischer, et al. [41] suggested employing the likelihood ratio to evaluate a test's

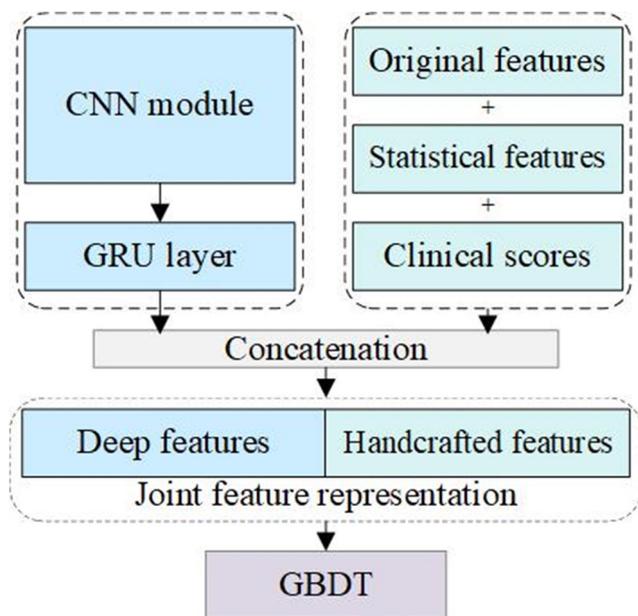


Fig. 4 The processing of early fusion

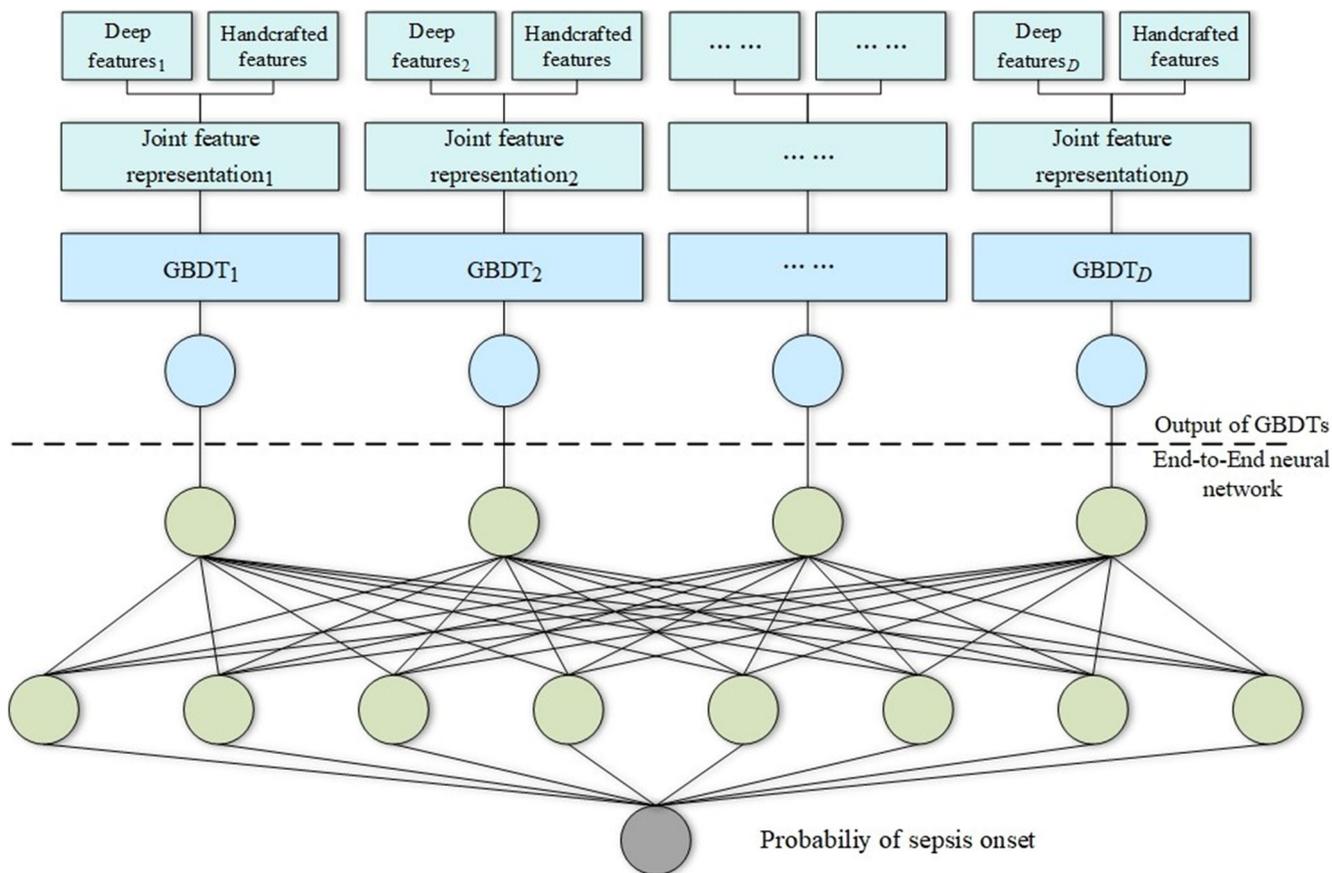


Fig. 5 The main process of the double fusion framework

predictive qualities. Equations (11) – (15) can be used to calculate accuracy, sensitivity, and specificity:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{12}$$

$$\text{specificity} = \frac{TN}{TN + FP} \tag{13}$$

$$LHR+ = \frac{\text{sensitivity}}{1 - \text{specificity}} \tag{14}$$

$$LHR- = \frac{1 - \text{sensitivity}}{\text{specificity}} \tag{15}$$

where TP, TN, FP, and FN are the number of true positive test samples, true negative test samples, false positive test samples, and false negative test samples, respectively. The model was implemented using Python programming language, on a personal computer with Windows 10 64-bit, and an NVIDIA RTX 3060Ti having a 3.80 GHz AMD Ryzen 7 5800X CPU with 48 GB of RAM.

3 Results

3.1 Data analysis

To clarify the details of the used dataset, this subsection first provides statistics and characteristics of the used dataset, followed by an illustrative example from the dataset. The statistical outcome is shown in Table 3. The sepsis prevalence in the overall cohort is 40%. The males account

Table 3 The statistics and characteristics of the hospital in Shanghai

| | Total | Sepsis | No-Sepsis |
|--------|-----------|----------|-----------|
| Total | 282(100%) | 114(40%) | 168(60%) |
| Gender | | | |
| Male | 147(52%) | 68(24%) | 79(28%) |
| Female | 135(48%) | 46(16%) | 89(32%) |
| Age | | | |
| 21–40 | 11(4%) | 2(1%) | 9(3%) |
| 41–60 | 23(8%) | 7(2%) | 16(6%) |
| 61–80 | 105(37%) | 47(17%) | 58(21%) |
| 81+ | 143(51%) | 58(21%) | 85(30%) |

for 52% of the cohort, while females account for 48%. The majority of patients in this dataset are over 60 years old, and their risk of developing sepsis is higher than that of patients under 40. The hospital in Shanghai dataset contains 38,884 h of recorded ICU data.

Table 4 shows an illustrative example of a patient's first 8 hours in the ICU. As shown in Table 4, the patient information including vital signs, laboratory test results, demographic, and sepsis results are given. In this case, the patient's age is 68 and his sepsis result is negative during the first 8 hours of ICU. Due to the irregular intervals and varying frequencies, there are some missing values in the laboratory test result, such as PaCO₂ and FiO₂. Furthermore, since not all laboratory test result of the patient is collected, there are some missing features, such as IL-6. Each data instance in this experiment consists of multiple data series. In total, the dataset used in the experiment contains 38,884 samples and 31 attributions, including input features with 10 vital signs, 17 laboratory test results, and 3 demographics, and a classification label with sepsis result.

3.2 An illustrative case of DFSP prediction

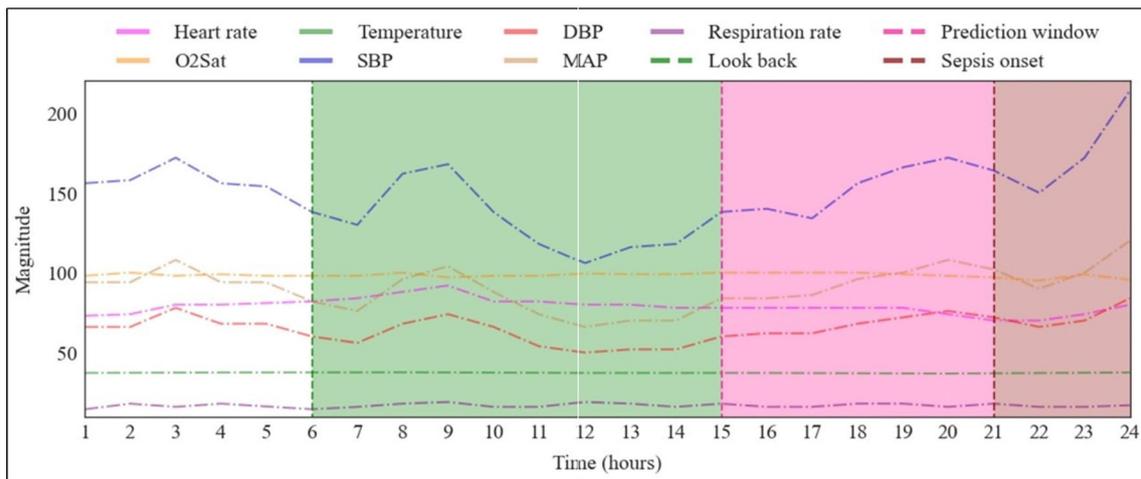
In this subsection, an illustrative case is used to express the prediction of DFSP, where a sepsis patient is randomly selected from the dataset for this. The prediction window is set at 6 h, and the length of the look back is set at 10 h. Figure 6(a) illustrates the patient's hourly vital signs, such as heart rate, temperature, SBP, DBP, MAP, O₂Sat, and respiratory rate. The hourly prediction risk scores of DFSP are shown in Fig. 6(b). As shown, the patient is diagnosed with sepsis in the 21st hour, while the risk score of DFSP exceeds the threshold level at the 15th hour and generates a sepsis warning, confirming that DFSP correctly predicts the onset of sepsis for this patient 6 hours in advance.

3.3 Classification result

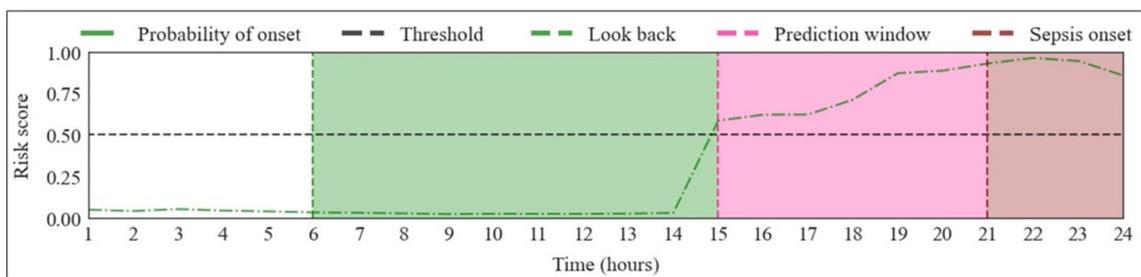
The proposed model is evaluated by examining its ability to predict sepsis 6 h, 12 h, and 24 h before it is developed. A 10-fold cross-validation scheme is performed, i.e., in each of the ten cross-validation runs, 90% of the patients are selected as training cohorts, with their patient records being used to build the model, while the records of remaining patients are selected for testing. DFSP's prediction performance is assessed by comparing two early-warning scores (SIRS and qSOFA), and three existing machine-learning approaches, as baselines: 1) Deep SOFA-Sepsis Prediction Algorithm (DSPA): Asuroglu and Ogul [42] developed the hybrid deep learning model that combines the CNN and random forest to predict SOFA scores of sepsis patients with a significant performance,

Table 4 An illustrative example of a patient's first 8 hours of ICU stay

| Demographic | | | Vital sign | | | | Laboratory test result | | | | | Sepsis result |
|-------------|-----|-----|------------------|------------------|------------------------|-----|--------------------------|----------|--------------|------------------|-----|---------------|
| ICULOS (h) | Sex | Age | Heart rate (BPM) | Temperature (°C) | Respiratory rate (BPM) | ... | PaCO ₂ (mmHg) | Blood ph | IL-6 (pg/mL) | FiO ₂ | ... | Sepsis result |
| 0 | Man | 68 | 108 | 36.3 | 20 ... | ... | 7.4 | - | - | ... | ... | Negative |
| 1 | Man | 68 | 110 | 36.3 | 20 | ... | 30 | - | - | - | ... | Negative |
| 2 | Man | 68 | 112 | 36.25 | 21 | ... | - | - | - | - | ... | Negative |
| 3 | Man | 68 | 114 | 36.23 | 22 | ... | 31 | - | - | 45 | ... | Negative |
| 4 | Man | 68 | 116 | 36.2 | 22 | ... | 31 | - | - | - | ... | Negative |
| 5 | Man | 68 | 115 | 36.28 | 22 | ... | - | - | - | - | ... | Negative |
| 6 | Man | 68 | 114 | 36.35 | 23 | ... | - | - | - | - | ... | Negative |
| 7 | Man | 68 | 113 | 36.43 | 23 | ... | 32 | - | - | - | ... | Negative |
| 8 | Man | 68 | 112 | 36.5 | 23 | ... | - | - | - | - | ... | Negative |



(a) The hourly vital signs of a sepsis patient



(b) The hourly risk score from DFSP for a sepsis patient

Fig. 6 An illustrative case of a sepsis patient

2) Residual network (ResNet): He, et al. [43] proposed this effective deep learning framework that addresses the vanishing/exploding gradient issue and succeeds at a variety of classification tasks, 3) Time-phAsed: Li, et al. [26] developed this XGBoost based method, which performs well in the 2019 PhysioNet/Computing in Cardiology Challenge dataset. To assess the performance of the above machine learning models, the accuracy, specificity, and AUC score at a fixed 0.80 sensitive level were calculated. The ROC curve comparison of DFSP and baseline models is given in Fig. 7, where the prediction window is set at 6 h. From Fig. 7, it is clear that DFSP proposed in this paper performs better than all the baselines.

Table 5 shows the accuracy, specificity, sensitivity, and AUC score of DFSP of this experiment, for prediction windows of 6 h, 12 h, and 24 h. The results show that DFSP has the highest AUC score across all prediction windows. When the prediction is performed over larger prediction windows, the AUC score of DFSP decreased from 0.92 at the 6 h prediction window to 0.89 at the 24 h prediction window. DFSP achieves a higher specificity of 0.87, with a sensitivity of 0.80 in a 6 h prediction window, compared with the Time-phAsed’s 0.81 specificity with a sensitivity of 0.80, the DSPA’s 0.73 specificity with a sensitivity of 0.80

and the ResNet’s 0.73 specificity with a sensitivity of 0.80. This result is consistent with the calculated AUC scores, which measure sensitivity and specificity. In addition, DFSP has much higher accuracy than baselines for 6 h, 12 h, and 24 h prediction windows.

Figure 8 shows the LHR+ and LHR- of DFSP of this experiment for prediction windows of 6 h, 12 h, and 24 h. According to Fig. 8, DFSP has the highest LHR+ and the lowest LHR- across all prediction windows, which clearly indicates that the DFSP outperforms baselines in terms of diagnosing performance.

3.4 Ablation study

To measure the superiority of the proposed double fusion framework, an ablation study was conducted, which removes the hybrid deep learning model, early fusion, late fusion, and double fusion in DFSP. Along with the evaluation of DFSP advantages in prediction performance, this experiment also assesses the improvements in robustness brought on by late fusion. To assess the robustness of the model, each model is performed 15 times 10-fold cross-validation to assess its stability. In addition, the performance

Fig. 7 ROC curves of DFSP and baselines using a 6 h prediction window

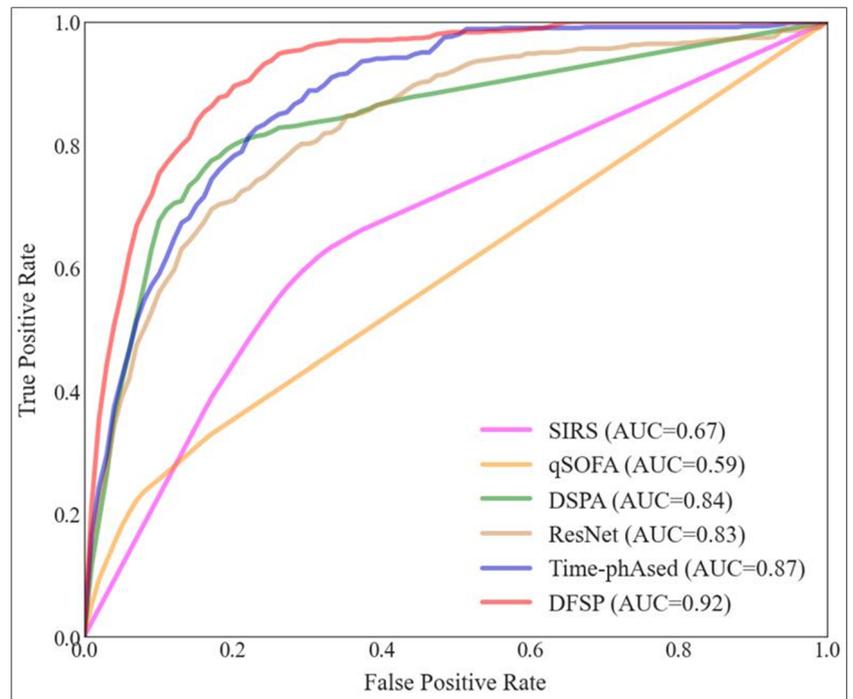
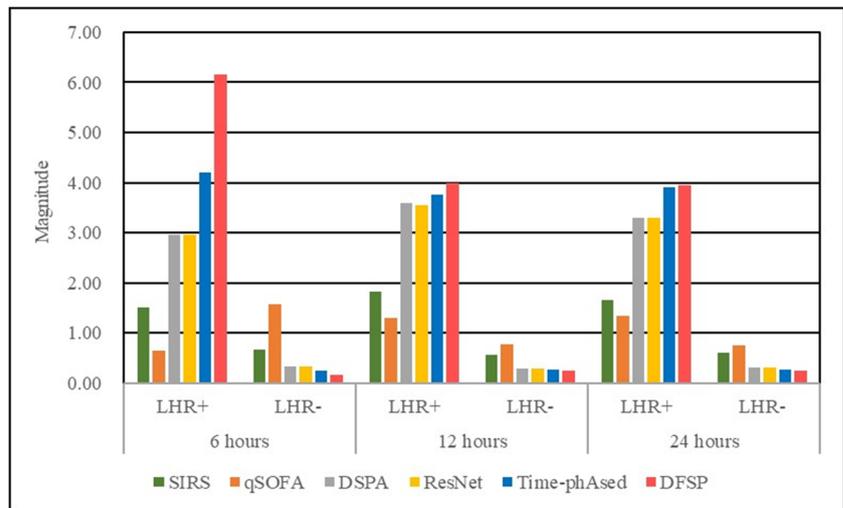


Table 5 The classification results of DFSP and baselines

| Method | AUC | Acc | Spec | Sens |
|------------------------|-------------|-------------|-------------|-------------|
| 6 h prediction window | | | | |
| SIRS | 0.67 | 0.57 | 0.57 | 0.65 |
| qSOFA | 0.59 | 0.55 | 0.56 | 0.28 |
| DSPA | 0.84 | 0.74 | 0.73 | 0.80 |
| ResNet | 0.83 | 0.73 | 0.73 | 0.80 |
| Time-phAsed | 0.87 | 0.81 | 0.81 | 0.80 |
| DFSP | 0.92 | 0.87 | 0.87 | 0.80 |
| 12 h prediction window | | | | |
| SIRS | 0.65 | 0.57 | 0.58 | 0.68 |
| qSOFA | 0.57 | 0.55 | 0.56 | 0.57 |
| DSPA | 0.83 | 0.71 | 0.72 | 0.80 |
| ResNet | 0.82 | 0.71 | 0.71 | 0.80 |
| Time-phAsed | 0.85 | 0.75 | 0.75 | 0.80 |
| DFSP | 0.89 | 0.80 | 0.80 | 0.80 |
| 24 h prediction window | | | | |
| SIRS | 0.63 | 0.58 | 0.58 | 0.65 |
| qSOFA | 0.56 | 0.36 | 0.39 | 0.71 |
| DSPA | 0.81 | 0.65 | 0.66 | 0.80 |
| ResNet | 0.81 | 0.65 | 0.66 | 0.80 |
| Time-phAsed | 0.85 | 0.78 | 0.78 | 0.80 |
| DFSP | 0.89 | 0.78 | 0.79 | 0.80 |

*Significant value is boldfaced

Fig. 8 LHR+ and LHR- of DFSP and baselines



of DFSP without the double fusion is presented, by testing the hybrid deep learning model. The results of the ablation study are shown in Table 6, which includes four statistical indicators with the best value, worst value, average value, and standard deviation in AUC, being denoted as “Best”, “Worst”, “Mean”, and “SD”, respectively. As Table 6 shows, DFSP performs better when using a hybrid deep learning model than when using other deep learning models, proving that the proposed hybrid deep learning model is highly capable of extracting useful information from EMRs data. The results also reveal that DFSP with fusion strategies can achieve a better predictive performance

than the hybrid deep learning model, demonstrating that the proposed fusion strategies are effective for improving performance. The results demonstrate that early and late fusions enhance the DFSP in various ways. The early fusion strategy is effective at improving predictive performance, and its incorporation significantly raises the AUC score of DFSP. The late fusion strategy is successful in enhancing stability, which can be evidenced by the fact that the standard deviations of DFSP and other models including late fusion are less than 0.005 and better than that of models without late fusion. The DFSP achieves the best performance among all variants of the model, which indicated that

Table 6 The results of the ablation study

| Method | AUC | | | |
|----------------------------|-------------|-------------|-------------|--------------|
| | Best | Worst | Mean | Std |
| 6 h prediction window | | | | |
| Hybrid deep learning model | 0.82 | 0.79 | 0.8 | 0.018 |
| DFSP without early fusion | 0.83 | 0.83 | 0.83 | 0.00 |
| DFSP without late fusion | 0.91 | 0.87 | 0.89 | 0.017 |
| DFSP | 0.92 | 0.91 | 0.92 | 0.004 |
| 12 h prediction window | | | | |
| Hybrid deep learning model | 0.80 | 0.75 | 0.78 | 0.015 |
| DFSP without early fusion | 0.81 | 0.8 | 0.81 | 0.004 |
| DFSP without late fusion | 0.88 | 0.85 | 0.87 | 0.011 |
| DFSP | 0.90 | 0.89 | 0.89 | 0.003 |
| 24 h prediction window | | | | |
| Hybrid deep learning model | 0.80 | 0.78 | 0.79 | 0.008 |
| DFSP without early fusion | 0.79 | 0.78 | 0.79 | 0.004 |
| DFSP without late fusion | 0.88 | 0.85 | 0.87 | 0.012 |
| DFSP | 0.89 | 0.88 | 0.89 | 0.003 |

*Significant value is boldfaced

the double fusion framework is successful in combining the benefits of early and late fusion.

3.5 Computational complexity analysis

To evaluate the computational complexity of DFSP, the computational time required to make predictions using DFSP and other machine learning models was measured. The prediction window was set at 6 h, and all patient data are used in this experiment. In Table 7 it is presented the computational time required for each model to make the prediction.

As shown in Table 7, the computational time of making predictions using DFSP is 5.35s, which is approximately the sum of D hybrid deep learning models, D GDBTs, and the End-to-End network. The results of the comparison demonstrate that DSPA and ResNet, both using the deep learning model, take approximately the same amount of time to compute as DFSP. When compared to Time-phAsed, DFSP requires more computational time to make predictions, since DFSP must extract temporal features that can improve prediction performance. It is reasonable to conclude that the computational complexity of DFSP is appropriate, as it can produce a significant prediction result in acceptable computational times.

4 Discussion

Accurate early prediction of sepsis can enhance septic patient survival and lower hospital costs. Thus, in this paper it is proposed an early sepsis predictor called DFSP, which is a double fusion framework that combines early fusion and late fusion. The major benefit of DFSP is the integration of auto-extracted knowledge from deep learning models and clinical knowledge, based on clinical experience. Aside from extracting additional information from several sources, the proposed double fusion framework is also capable of solving the missing value and high

sparsity problem, which is inevitable in the clinical area. First, depending on the patient's medical condition, doctors may only need to detect a subset of clinical features of interest, which means that not all clinical features are observed in each patient. Second, as some clinical features have longer data-collection time intervals than others, some time series features have a high sparsity. Missing value and high sparsity problem typically lead to consequences such as loss of algorithmic performance and sample bias [44]. Several studies have demonstrated that this has a negative effect on forecasting and classification tasks [45, 46]. There were usually two approaches to dealing with missing and sparse data: one is to use a deep learning model with interpolation strategies, while the other is the tree-based model with handcrafted features. Nevertheless, the former generates unexpected noise and performs poorly in sparse time series. The latter typically compacts information granules, resulting in significant information loss. Besides, the previous study has also shown that it suppresses critical fine-grained information since the granularity of observed time series might vary from patient to patient depending on the underlying medical condition [47]. DFSP overcomes the disadvantages of the deep learning model and tree-based model, by fusing deep and handcrafted features. On the one hand, using deep features can help avoid information loss caused by handcrafted features. Data compacted through handcrafted features, on the other hand, can remove data noise and missing problems.

To evaluate the framework, a retrospective study was conducted, by using patients admitted to the ICUs of a hospital in Shanghai. For a 6 h prediction window, DFSP had an AUC score of 0.92 and achieved a specificity of 0.87 and an accuracy of 0.87 with a sensitivity of 0.80. Likelihood ratios were used as an alternative measure to predictive value, since it is less influenced by the prevalence of sepsis in the population. At a 6 h prediction window, DFSP can provide 6.15 of LHR+ and 0.16 of LHR-. Additionally, the prediction windows were varied to test the performance of DFSP. The proposed model was compared with two existing early-warning scores and three state-of-the-art methods for sepsis prediction, having been demonstrated that the DFSP outperforms the baselines.

However, this study has a few limitations. To begin with, DFSP can fuse information from unstructured data to improve its performance, but it was not possible to evaluate it because the used dataset lacks unstructured data such as radiological images and clinical notes. Secondly, DFSP needs to be retrained for the other datasets since the different datasets have different clinical measures and sepsis definitions. In practice, a well-trained sepsis prediction model should be fed data from multiple data sources. This difficulty has also been identified in previous studies of

Table 7 The computational time required for each model to make the prediction

| Model | Time (s) |
|----------------------------|----------|
| DFSP | 5.35 |
| Hybrid deep learning model | 1.02 |
| GDBT | 0.29 |
| End-to-End network | 0.05 |
| DSPA | 4.42 |
| ResNet | 5.40 |
| Time-phAsed | 0.28 |

sepsis prediction; thus, it can be considered as one of future research directions, such as using transfer learning to make the DFSP more practical.

Future research can investigate whether the prediction performance of the model can be further improved, by fusing some strong feature representations from different deep learning modules to extract. Furthermore, since various medical events may be interconnected, there may be interdependence among features. To enhance the predictive performance of models for sepsis prediction, the proposal of a more effective feature selection method is needed to improve the DFSP, which can help to select the best set of discriminatory features from the fused feature representation.

Another potential path for this paper is to improve the DFSP's earliness. The earliness of sepsis prediction is crucial for improving sepsis management. Thus, the objective of sepsis prediction should be to identify the sepsis as soon as possible, while maintaining prediction accuracy. The experiment results revealed that when the prediction window got larger, the performance of DFSP and most machine learning models decreases, which is consistent with the result in Scherpf, et al. [20], Rafiei, et al. [48], and Shashikumar, et al. [23], where the performance of sepsis prediction models degraded as prediction window grew larger. It may be due to the fact that several significant signs for diagnosing sepsis, such as shortness of breath, high fever, and abnormal heart rate [2], usually appear near the onset of sepsis. Furthermore, several recent studies on time series prediction problems show that the accuracy and earliness of models are frequently in conflict [49]. As a result, the sepsis prediction problem should be regarded as a multi-objective optimization problem, which it is intended to tackle in the future with the evolutionary algorithm.

Acknowledgements The work has been supported by National Natural Science Foundation of China (No. 72171176, NO. 71771179, NO. 72021002) and China-Germany Cooperation project supported by National Natural Science Foundation of China (M-310).

Data availability The data that support the findings of this study are available from the authors upon reasonable request.

Author Contributions Yongrui Duan and Mingzhou Chen conceived and designed this study. Yongrui Duan, Mingzhou Chen, Fenggang Hou, Guoliang Yan, Shufang Li, and Haihui Wang performed the experiments. Yongrui Duan, Mingzhou Chen and Jiazheng Huo assisted with research and writing. All authors were responsible for data interpretation and statistical analysis. All authors assisted in drafting and revising the manuscript at all stages, and all have approved it in this final form.

Declarations

Conflict of Interests The authors declare that there is no conflict of interests regarding the publication of this article.

Ethical standard The authors state that this research complies with ethical standards. This research does not involve either human participants or animals.

References

1. Leone M (2016) Septic shock: a global response. *Presse Med* 45(4):E91. <https://doi.org/10.1016/j.lpm.2016.03.002>
2. Jaimes F, Garces J, Cuervo J, Ramirez F, Ramirez J, Vargas A, Quintero C, Ochoa J, Tandioy F, Zapata L, Estrada J, Yepes M, Leal H (2003) *Intensive Care Med* 29(8):1368. <https://doi.org/10.1007/s00134-003-1874-0>
3. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, Cohen J, Opal SM, Vincent JL, Ramsay G (2003) C Int sepsis definitions. *Crit Care Med* 31(4):1250. <https://doi.org/10.1097/01.ccm.0000050454.01978.3b>
4. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Cooper-Smith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, van der Poll T, Vincent JL, Angus DC (2016) *Jama-J Amer Med Assoc* 315(8):801. <https://doi.org/10.1001/jama.2016.0287>
5. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna T, Kadri SS, Angus DC, Danner RL, Fiore AE, Jernigan JA, Martin GS, Septimus E, Warren DK, Karcz A, Chan C, Menchaca JT, Wang R, Gruber S, Klompas M, Program CDCPE (2017) *Jama-J Amer Med Assoc* 318(13):1241. <https://doi.org/10.1001/jama.2017.13836>
6. Liang L, Moore B, Soni A (2020) National inpatient hospital costs: The most expensive conditions by payer, 2017: Statistical brief #261 (Agency for healthcare research and quality (US), Rockville (MD) 14 Jul 2020)
7. Morrill JH, Kormilitzin A, Nevado-Holgado AJ, Swaminathan S, Howison SD, Lyons TJ (2020) *Crit Care Med* 48(10):E976. <https://doi.org/10.1097/ccm.0000000000004510>
8. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna T, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM (2017) *N Engl J Med* 376(23):2235. <https://doi.org/10.1056/NEJMoa1703058>
9. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, Machado FR, McIntyre L, Ostermann M, Prescott HC, Schorr C, Simpson S, Wiersinga WJ, Alshamsi F, Angus DC, Arabi Y, Azevedo L, Beale R, Beilman G, Belley-Cote E, Burry L, Cecconi M, Centofanti J, Coz Yataco A, De Waele J, Dellinger RP, Doi K, Du B, Estenssoro E, Ferrer R, Gomersall C, Hodgson C, Moller MH, 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, Papanthanasoglou E, Perner A, Puskarich M, Roberts J, Schweickert W, Seckel M, Sevransky J, Sprung CL, Welte T, Zimmerman J, Levy M (2021) *Intensive Care Med*. <https://doi.org/10.1007/s00134-021-06506-y>
10. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D, Kumar A, Cheang M (2006) *Crit Care Med* 34(6):1589. <https://doi.org/10.1097/01.ccm.0000217961.75225.e9>
11. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JLL, Tan GYH (2021) *Nature Commun*, vol 12(1). <https://doi.org/10.1038/s41467-021-20910-4>
12. Lever A, Mackenzie I (2007) *Bmj-British Med J* 335(7625):879. <https://doi.org/10.1136/bmj.39346.495880.AE>
13. Lim J, Lee YY, Choy YB, Park W, Park CG (2021) *Biomed Eng Lett* 11(3):197. <https://doi.org/10.1007/s13534-021-00200-0>

14. Calvert J, Price DA, Chettipally U, Barton C, Feldman MD, Hoffman J, Jay M, Das R (2016) *Comput Biol Med* 74:69. <https://doi.org/10.1016/j.combiomed.2016.05.003>
15. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ, Das R (2016) *Jmir Med Inf* 4(3):67. <https://doi.org/10.2196/medinform.5909>
16. Mao Q, Jay M, Hoffman J, Calvert J, Barton C, Shimabukuro D, Shieh L, Chettipally U, Fletcher G, Kerem Y, Zhou Y, Das R (2018) *bmj open*, vol 8(1). <https://doi.org/10.1136/bmjopen-2017-017833>
17. Kok C, Jahmunah V, Oh SL, Zhou X, Gururajan R, Tao X, Cheong KH, Gururajan R, Molinari F, Acharya UR (2020) *Comput Bio Med*, vol 127. <https://doi.org/10.1016/j.combiomed.2020.103957>
18. Zhang D, Yin C, Hunold KM, Jiang X, Caterino JM, Zhang P (2021) An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns (New York, N.Y.)* 2(2):100196. <https://doi.org/10.1016/j.patter.2020.100196>
19. Wang Z, Yao B (2022) *Ieee J Biomed Health Inf* 26(2):876. <https://doi.org/10.1109/jbhi.2021.3092835>
20. Scherpf M, Graesser F, Malberg H, Zaunseder S (2019) Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput Biol Med*, vol 113. <https://doi.org/10.1016/j.combiomed.2019.103395>
21. Kam HJ, Kim HY (2017) Semantic content analysis and annotation of histological images. *Comput Biol Med* 89:248. <https://doi.org/10.1016/j.combiomed.2017.08.015>
22. Fagerstrom J, Bang M, Wilhelms D, Chew MS (2019) *Scientific reports*, vol 9. <https://doi.org/10.1038/s41598-019-51219-4>
23. Shashikumar SP, Josef CS, Sharma A, Nemati S (2021) DeepAISE - An interpretable and recurrent neural survival model for early prediction of sepsis. *Artif Intell Med*, vol 113. <https://doi.org/10.1016/j.artmed.2021.102036>
24. Wang D, Li J, Sun Y, Ding X, Zhang X, Liu S, Han B, Wang H, Duan X, Sun T (2021) A machine learning model for accurate prediction of sepsis in ICU patients. *Frontiers Public Health*, vol 9. <https://doi.org/10.3389/fpubh.2021.754348>
25. Barton C, Chettipally U, Zhou Y, Jiang Z, Lynn-Palevsky A, Le S, Calvert J, Das R (2019) Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Comput Biol Med* 109:79. <https://doi.org/10.1016/j.combiomed.2019.04.027>
26. Li X, Xu X, Xie F, Xu X, Sun Y, Liu X, Jia X, Kang Y, Xie L, Wang F, Xie G (2020) A time-phased machine learning model for real-time prediction of sepsis in critical care. *Crit Care Med* 48(10):E884. <https://doi.org/10.1097/ccm.0000000000004494>
27. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput Vis Image Understand* 156:117
28. Sun D, Wang M, Li A (2018) A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans Computat Biol Bioinform* 16(3):841
29. Lv J, Hu X, Li L, Li P (2019) An effective confidence-based early classification of time series. *IEEE Access* 7:96113. <https://doi.org/10.1109/ACCESS.2019.2929644>
30. Hagerty JR, Stanley RJ, Almubarak HA, Lama N, Kasmir R, Guo P, Drugge RJ, Rabinovitz HS, Oliviero M, Stoeker WV (2019) Deep learning and handcrafted method fusion: higher diagnostic accuracy for melanoma dermoscopy images. *IEEE J Biomed Health Inf* 23(4):1385
31. Zhang Y, Chen W (2022) Decision-level information fusion powered human pose estimation. *Appl Intell*. <https://doi.org/10.1007/s10489-022-03623-z>
32. Ilhan HO, Serbes G, Aydin N (2022) Decision and feature level fusion of deep features extracted from public COVID-19 data-sets. *Appl Intell* 52(8):8551. <https://doi.org/10.1007/s10489-021-02945-8>
33. Zuo Q, Zhang J, Yang Y (2021) DMC-fusion: deep multi-cascade fusion with classifier-based feature synthesis for medical multi-modal images. *Ieee J Biomed Health Inf* 25(9):3438. <https://doi.org/10.1109/jbhi.2021.3083752>
34. Gumaei A, Ismail WN, Hassan MR, Hassan MM, Mohamed E, Alelaiwi A, Fortino G (2022) A decision-level fusion method for COVID-19 patient health prediction. *Big Data Res*, vol 27. <https://doi.org/10.1016/j.bdr.2021.100287>
35. Wang H, Hu J, Song Y, Zhang L, Bai S, Yi Z (2022) Multi-view fusion segmentation for brain glioma on CT images. *Appl Intell* 52(7):7890. <https://doi.org/10.1007/s10489-021-02784-7>
36. Dey R, Salem FM (2017) *Ieee*. In: 60th IEEE international midwest symposium on circuits and systems (MWSCAS). Midwest symposium on circuits and systems conference proceedings, pp 1597–1600
37. Kingma D, Ba J (2014) *Computer science*
38. Zhang Y, Sidibe D, Morel O, Meriaudeau F (2021) Deep multi-modal fusion for semantic image segmentation: a survey. *Image Vis Comput*, vol 105. <https://doi.org/10.1016/j.imavis.2020.104042>
39. Boulahia SY, Amamra A, Madi MR, Daikh S (2021) Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Mach Vis Appl* 32(6):1
40. Deeks JJ, Altman DG (2004) Diagnostic tests 4: likelihood ratios. *Br Med J* 329(7458):168. <https://doi.org/10.1136/bmj.329.7458.168>
41. Fischer JE, Bachmann LM, Jaeschke R (2003) A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med* 29(7):1043. <https://doi.org/10.1007/s00134-003-1761-8>
42. Asuroglu T, Ogul H (2021) *Computer methods and programs in biomedicine*, vol 198. <https://doi.org/10.1016/j.cmpb.2020.105816>
43. He K, Zhang X, Ren S, Sun J (2016) *Ieee*. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). IEEE conference on computer vision and pattern recognition, pp 770–778. <https://doi.org/10.1109/cvpr.2016.90>
44. Lin J, Li N, Alam MA, Ma Y (2020) Data-driven missing data imputation in cluster monitoring system based on deep neural network. *Appl Intell* 50(3):860. <https://doi.org/10.1007/s10489-019-01560-y>
45. Kwak SK, Kim JH (2017) Statistical data preparation: management of missing values and outliers. *Korean J Anesthesiology* 70(4):407. <https://doi.org/10.4097/kjae.2017.70.4.407>
46. Lu Y, Yang D, Li Z, Peng X, Zhong W (2022) Knowledge-based systems, vol 243. <https://doi.org/10.1016/j.knosys.2022.108510>
47. Tipirneni S, Reddy CK (2022) *Acm transactions on knowledge discovery from data*, vol 16(6). <https://doi.org/10.1145/3516367>
48. Raffei A, Rezaee A, Hajati F, Gheisari S, Golzan M (2021) SSP: early prediction of sepsis using fully connected LSTM-CNN model. *Comput Biol Med*, vol 128. <https://doi.org/10.1016/j.combiomed.2020.104110>
49. Mori U, Mendiburu A, Dasgupta S, Lozano JA (2018) Early classification of time series by simultaneously optimizing the accuracy and earliness. *Ieee Trans Neural Netw Learn Syst* 29(10):4569. <https://doi.org/10.1109/tnnls.2017.2764939>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Affiliations

Yongrui Duan¹ · Jiazhen Huo¹ · Mingzhou Chen¹ · Fenggang Hou² · Guoliang Yan³ · Shufang Li⁴ · Haihui Wang³

Yongrui Duan
yrduan@tongji.edu.cn

Jiazhen Huo
huojiazhen@163.com

Fenggang Hou
fghou@qq.com

Guoliang Yan
13761788439@139.com

Shufang Li
13641650746@126.com

Haihui Wang
wanghaihui1982@126.com

- ¹ School of Economics & Management, Tongji University, Shanghai, China
- ² Department of Oncology, Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai, China
- ³ Department of Geriatrics, Shanghai Municipal Hospital of Traditional Chinese Medicine, Shanghai, China
- ⁴ Emergency Department, Shuguang Hospital Affiliated to Shanghai University of Traditional Chinese Medicine, Shanghai, China