RESEARCH



Inter-organ correlation based multi-task deep learning model for dynamically predicting functional deterioration in multiple organ systems of ICU patients



Zhixuan Zeng¹⁺, Yang Liu³⁺, Shuo Yao¹, Minjie Lin², Xu Cai¹, Wenbin Nan¹, Yiyang Xie¹ and Xun Gong^{1*}

⁺Zhixuan Zeng and Yang Liu are co-first authors.

*Correspondence: Xun Gong gongxun1246@csu.edu.cn ¹Department of Emergency Medicine, The Second Xiangya Hospital of Central South University, Changsha, China ²Academic Affairs Department, The Second Xiangya Hospital of Central South University, Changsha, China ³Department of Rehabilitation, The Second Xiangya Hospital of Central South University, Changsha, China

Abstract

Background Functional deterioration (FD) of various organ systems is the major cause of death in ICU patients, but few studies propose effective multi-task (MT) model to predict FD of multiple organs simultaneously. This study propose a MT deep learning model named inter-organ correlation based multi-task model (IOC-MT), to dynamically predict FD in six organ systems.

Methods Three public ICU databases were used for model training and validation. The IOC-MT was designed based on the routine MT deep learning framework, but it used a Graph Attention Networks (GAT) module to capture inter-organ correlation and an adaptive adjustment mechanism (AAM) to adjust prediction. We compared the IOC-MT to five single-task (ST) baseline models, including three deep models (LSTM-ST, GRU-ST, Transformer-ST) and two machine learning models (GRU-ST, RF-ST), and performed ablation study to assess the contribution of important components in IOC-MT. Model discrimination was evaluated by AUROC and AUPRC, and model calibration was assessed by the calibration curve. The attention weight and adjustment coefficient were analyzed at both overall and individual level to show the AAM of IOC-MT.

Results The IOC-MT had comparable discrimination and calibration to LSTM-ST, GRU-ST and Transformer-ST for most organs under different gap windows in the internal and external validation, and obviously outperformed GRU-ST, RF-ST. The ablation study showed that the GAT, AAM and missing indicator could improve the overall performance of the model. Furthermore, the inter-organ correlation and prediction adjustment of IOC-MT were intuitive and comprehensible, and also had biological plausibility.

Conclusions The IOC-MT is a promising MT model for dynamically predicting FD in six organ systems. It can capture inter-organ correlation and adjust the prediction for one organ based on aggregated information from the other organs.

Keywords Organ dysfunction, Multi-task learning, Deep learning, Gated recurrent unit, Graph attention networks, Multivariate time series



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licens es/by-nc-nd/4.0/.

Introduction

Organ dysfunction is commonly seen in critically ill patients and is the major cause of death in ICU [1, 2]. Patients may already have organ dysfunction at admission and then suffer a further functional deterioration (FD), or develop organ dysfunction during ICU stay. Furthermore, in most cases patients suffer FD in more than one organ system, which is well known as the multiple organ dysfunction syndrome (MODS) [3]. Due to the high mortality and the increased burden of health-care resource brought by FD [4, 5], early prediction of FD to initiate intervention is of important clinical significance.

In recent years, machine learning and/or deep learning models are increasingly adopted to predict FD and achieve the state-of-the-art predictive performance. Many previous studies developed their models for one single organ system, such as predicting the risk of acute kidney injury (AKI) [6], circulatory failure [7] or respiratory failure [8]. Some other studies aimed to predict MODS [9, 10], but without analyzing which organ system contributed to the MODS. Besides, there are also some studies focused on predicting multiple complications including FD in one or more organ systems [11, 12], but they trained separate models for predicting each complication respectively. All these studies essentially belong to single-task (ST) learning paradigm and their models are unable to serve as multi-organ warning system.

Multi-task (MT) learning aims to develop one model for handling multiple tasks simultaneously. Deep learning has become the priority scheme for MT due to its excellent ability of feature extraction [13]. The most universal MT deep learning framework is composed of a shared neural-network encoder and multiple task-specific output heads. The shared encoder extracts shared feature (high-dimensional vector) for all tasks, and then the output heads, which are generally composed of several linear layers, take in the shared feature to produce output for corresponding tasks respectively [13, 14]. Several previous studies proposed their models for multiple clinical tasks based on this MT framework but using different shared encoders. Harutyunyan H et al. [15] adopted long short-term memory (LSTM) network as the shared encoder to simultaneously predict hospital mortality, disease decompensation (real-time mortality), length of stay and disease phenotype. Cherifa M et al. [16] used gated recurrent unit (GRU) encoder to dynamically predict mean arterial pressure (MAP) and heart rate (HR). Roy S et al. [17] proposed a modified recurrent neural network (RNN) encoder called sequential subnetwork routing (SeqSNR) which can learn to use different subnetworks of the whole network for predicting dysfunction of different organs. These MT frameworks have simple architecture and greatly save computing resource compared to using multiple ST models, but a notable limitation is that their separate output heads are unable to model the potential correlation between different tasks. This disadvantage may lead to sub-optimal results since different organic systems in human body are closely connected and FD in one organ can affect another [18–20].

In this study, we propose an improved MT deep learning framework, named interorgan correlation based multi-task model (IOC-MT) for hourly dynamic prediction of the FD risk for six organ systems. The IOC-MT uses a Graph Attention Networks [21] (GAT) module to capture inter-organ correlation, and an adaptive adjustment mechanism (AAM) to adjust its prediction for an organ based on aggregated information from the other organs. Our experiment shows that the IOC-MT has comparable performance to the ST deep models, and better performance than the routine MT framework. In addition, we use the attention weight to show the inter-organ correlation and the adjustment coefficient to show the influence of inter-organ information aggregation on model output.

Method

Data source, participants and data extraction

We implemented a retrospective study on multivariate time series (MTS) data of patients from three public ICU databases: the Medical Information Mart for Intensive Care III (MIMIC-III) [22], MIMIC-IV [23] and eICU Collaborative Research Database (eICU-CRD) [24]. The MIMIC-III recorded patients admitted to ICUs of the Beth Israel Deaconess Medical Center between 2001 and 2012, while the MIMIC-IV recorded patients in this hospital between 2008 and 2019. The eICU-CRD recorded patients admitted to 335 ICUs of 208 hospitals located throughout the US during 2014 to 2015. A local ethical review board (ERB) approval was achieved for building these databases, thus a ERB approval from our institution was exempted.

We selected participants from the three databases according to the following inclusion criteria: (1) aged between 16 and 89 years old; (2) the first ICU stay of a patient; (3) length of ICU stay not less than 48 h; (4) the admission time is between 2014 and 2019 (for MIMIC-IV). As the MIMIC-IV recorded the actual time range of ICU admission as: 2008–2010, 2011–2013, 2014–2016 or 2017–2019, we used the above fourth inclusion criterion on MIMIC-IV to avoid time overlapping with MIMIC-III that might cause repeated inclusion of the same patients.

We extracted MTS data at hourly resolution for 38 dynamic clinical variables (Appendix A.1). Firstly, we divided each ICU stay into a sequence of continuous and non-overlapping hourly intervals. Secondly, we collected the measurements of the 38 variables during every interval. Thirdly, in each interval, if there was only one measurement for a variable, we used this measurement for representation; if there were multiple measurements for a variable, we used their aggregated measurement (maximum, minimum or mean as appropriate) (Appendix A.1); if there was no measurement for a variable, we marked it as missingness. After that we obtained the MTS of each ICU stay as our model input. We also extracted the static demographic data such as age, gender, admission type for statistical analysis, but not as sequential input of our models. Besides, for very few ICU stays with excessive length of hospitalization, we used their time series within 14 days after admission to ICU in model training or validation.

Notations and task statement

We denoted an MTS of D clinical variables with a length of L hours as $X = (x_1, \ldots, x_L)^{\mathrm{T}} \in \mathbf{R}^{L \times D}$, where $x_t = (x_t^1, \ldots, x_t^D) \in \mathbf{R}^D$ was a vector of the D variables in the *t*-th hour. $x_t^d \in \mathbf{R}$ was the measurement or aggregated measurement of the *d*-th variable in the *t*-th hour, and $x_t^d = null$ if it was missing. We additionally introduced a matrix $M = (m_1, \ldots, m_L)^{\mathrm{T}} \in \{0, 1\}^{L \times D}$ to indicate data missingness in X, where $m_t^d = 0$ indicated that x_t^d is missing, while $m_t^d = 1$ otherwise. Then the ICU stay of each included patient was denoted as a sample S = (X, M).

In this study we focused on multi-task prediction of FD in the six organ systems proposed by the Sequential Organ Failure Assessment (SOFA) score [25]. We chose SOFA score as it is the most widely used criteria to quantify organ dysfunction [2, 25, 26] and all the components in SOFA are available in the three ICU databases. In order to perform prediction at an hourly frequency, we assessed the SOFA score for each organ system in every hourly interval along the time series. The specific assessment rule was as: (1) if there was newly observed measurement related to an organ in current hour, the SOFA score of this organ was updated; (2) if no related measurement was observed in current hour, the SOFA score in previous hour was used; (3) if no related measurement was observed in the first hour of an ICU stay, the corresponding SOFA score was set to zero (defaulted as normal). In addition, for central nervous system (CNS) scoring, we defaulted the verbal Glasgow score (GCS) to be five if the patient had tracheal intubation, and for renal system scoring, we disused the 24-hour urine-output criterion in the first 23 h of an ICU stay as it was not available.

After assessing hourly SOFA scores, we defined hourly binary labels for the six organ systems (i.e. deteriorating = 1 and not-deteriorating = 0). We adopted the time-window setting similar to two previous studies [10, 16] which set an observation window, a gap window and a prediction window (Fig. 1). Specifically, when the model made prediction at the *t*-th hour, the observation window was the period from admission to the *t*-th hour and the MTS data during this period was analyzed for prediction; the prediction window was a future period used to identify whether FD occurred, and the label of an organ system was defined as deteriorating at the *t*-th hour if the maximum SOFA score of this organ system in the prediction window rose \geq 1 score compared to that in the *t*-th hour [2]; gap window was the period between the observation and prediction window and it is preserved for conducting clinical intervention in advance. In this study we set the prediction window to be 4 h, and the gap window to be 4, 8, 16 and 24 h respectively to develop and validate our models. For a series of *L* hours, we performed hourly prediction until the (L - gap window - prediction window)-th hour as labels of the following hours were indeterminable under our time-window setting.

Proposed model

Our proposed IOC-MT contained a shared encoder and six task-specific (or to say organ-specific) output heads which were correlated by the GAT module (Fig. 2). In this section we respectively introduced the shared encoder, the correlated output heads, the AAM and the loss function for model training.

Shared encoder

The shared encoder was composed of a linear layer (*Linear_enc*) and a GRU network. It encoded an MTS sample S = (X, M) to a sequence of shared features $F_s \in \mathbf{R}^{L \times D_s}$ as following:

$$F_s = GRU(Linear_enc(X || M)) \tag{1}$$



Fig. 1 Time axis for dynamic prediction (gap window = 8 h)



Fig. 2 Overall architecture of IOC-MT (it only demonstrated the computation of the first output head for brevity)

where symbol || denoted concatenation of two matrices to produce $(X||M) \in \mathbf{R}^{L \times 2D}$. D_s is the dimensionality of shared feature and the $Linear_enc$ also had D_s neurons.

Correlated output heads

Each organ-specific output head contained two linear layers followed by a sigmoid activation function, and a GAT module was used between the first and second linear layers of all the six output heads for information aggregation.

The first linear layer was responsible for projecting the F_s to organ-specific features. We used the following indices to denote different organ systems: {1: *Respiration*; 2: *Coagulation*; 3: *Liver*; 4: *Cardiovascular*; 5: *CNS*; 6: *Renal*}, and the organ-specific features for the *i*-th organ F_i were computed as:

$$F_i = \text{Linear 1}_i(F_s), \text{ where } i \in \{1, 2, 3, 4, 5, 6\}$$
 (2)

where $Linear1_i$ was the first linear layer of the *i*-th output head and all the six first linear layers had D_s neurons, so F_i had the same size of $L \times D_s$ as F_s .

Then the GAT module took these organ-specific features as input and output corresponding aggregated features. Specifically, in the *t*-th hour, we had six organ-specific features $F_i^t \in \mathbf{R}^{D_s}$, i = 1, ..., 6. The GAT constructed a graph of six nodes, where each node represented an organ system and F_i^t was node's feature. We illustrated the graphic structure and the adjacency matrix $A \in \{0, 1\}^{6 \times 6}$ of our GAT module in Fig. 2. The element in the *i*-th row and *j*-th column of A was denoted as a_{ij} , and $a_{ij} = 1$ indicated that there was an edge from node *i* to *j*, and $a_{ij} = 0$ otherwise. We constructed such a graph based on the clinical prior knowledge that each organ system may be affected by

Page 6 of 22

any other organ system. In this graph, given any two different nodes i and j there were two edges to model their correlation: e_{ij} and e_{ji} , where e_{ij} was the edge from i to jand e_{ji} was reverse. The value of e_{ij} indicated the importance of node i's feature F_i^t to node j. The GAT computed the value of each edge based on attention mechanism [21]. We used the scaled dot product attention [27] instead of the original attention mechanism of GAT as we found that it performed better. Taking node j's information aggregation as example, we firstly computed the values of the edges to node j as:

$$e_{ij} = \frac{\left(F_j^t W_q\right) \left(F_i^t W_k\right)^{\mathrm{T}}}{\sqrt{D_s}} \tag{3}$$

where *i* could be any node but except *j* itself, and W_q , $W_k \in \mathbf{R}^{D_s \times D_s}$ were the trainable parameters for query and key in attention mechanism. The e_{ij} was a scalar. Then we normalized the edge values into attention weights as:

$$\alpha_{ij} = softmax_i (e_{ij}) = \frac{exp(e_{ij})}{\sum_{1 \le k \le 6, k \ne j} exp(e_{kj})}$$
(4)

Using these attention weights, we computed weighted sum of the original features of node j's neighbor nodes as the aggregated feature of node j:

$$\widetilde{F}_{j}^{t} = \sum_{1 \le k \le 6, k \ne j} \alpha_{kj} F_{k}^{t}$$
(5)

Therefore the aggregated information from node, *j*'s neighbor nodes. After information aggregation, we subsequently used aggregated \widetilde{F}_{j} to adjust the original F_{j}^{t} as following:

$$\widehat{F}_{j}^{t} = (1 - \beta) F_{j}^{t} + \beta \widetilde{F}_{j}^{t}$$
(6)

where \widehat{F}_{j}^{t} was the adjusted feature of node j, and $\beta \in [0,1]$ was the coefficient determining adjustment strength. In order to make such an adjustment to be adaptive when time and organ varied, we let our model to learn the β rather than set β as a fixed hyperparameter. The β for node j at the *t*-th hour was computed as:

$$\beta_{j}^{t} = sigmoid(Linear_{\beta}(F_{j}^{t}||\widetilde{F}_{j}^{t}))$$
(7)

Where $Linear_{\beta}$ was a linear layer projecting $2D_s$ -dimension vector to scalar and it was identical across time series of all the six organ systems. Finally, the output was the predicted risk of FD in organ j at the *t*-th, which was:

$$Y_j^t = sigmoid(Linear2_j(\hat{F}_j^t))$$
(8)

where $Linear2_j$ was the second linear layer of the *j*-th output head. Notably, the whole IOC-MT performed parallel computation in all the output heads to output six predicted risks simultaneously.

Adaptive adjustment mechanism

As the above computation showed, the IOC-MT was able to adjust its original prediction for an organ after integrating related information from the other organs. As formula (5) showed, the larger the inter-organ attention weight α_{kj} was, the more important the organ k's original feature was for organ j. And as formula (6) showed, the larger the β value was, the more adjustment was made by the aggregated feature. Thus, we could clearly understand the AAM of IOC-MT through the α and β value. In this study, we analyzed the α and β value both at overall and individual level. At overall level, for each organ system, we counted all the β values over the time series of all the patients in external validation and calculated the sample proportion (here a sample referred to a certain hour in a patient's time series) in the β groups of 0.0-0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, 0.8-1.0. Then we calculated the average attention weight of samples in each β group to show the overall α assignment. At individual level, we showed the adjusted predicted risks and the original predicted risks of the IOC-MT for the six organ systems of a selected patient. The adjusted predicted risks were obtained by formula (8), while the original predicted risks were obtained by directly feeding the original organ-specific feature F_{j}^{t} into the corresponding second linear layer $Linear2_{j}$, without implementing GAT's information aggregation and adjustment. In addition, the α and β value and SOFA score for every organ at every hour were also showed for this patient. We used this individual example to demonstrate how the AAM worked.

Model training

The IOC-MT was trained by a joint binary cross entropy (BCE) loss function. Corresponding to Y_j^t in formula (8), we used $\overline{Y_j^t} \in \{0,1\}$ to represent the label of organ j at the *t*-th hour. Then the joint BCE loss for a sample of *L*-hour length was as:

$$Loss = \frac{1}{L - W} \sum_{t=1}^{L - W} \sum_{t=1}^{6} \bar{Y_j^t} \log(Y_j^t) + (1 - \bar{Y_j^t}) \log(1 - Y_j^t))$$
(9)

where $W = gap_window + prediction_window$, so L - W was the length of hourly prediction as mentioned before. The μ_j was the weight coefficient of organ j in the total loss. We set all μ_j to be 1/6 in this study, indicating that the six organ systems were equally important. The formula (9) was the joint loss for one sample, while for a batch of samples we used the mean of their joint losses.

Baseline models

We compared the IOC-MT to three ST deep sequential models and two ST nonsequential machine learning models as the following:

GRU single-task model (GRU-ST) It used the same *Linear*_*enc* and GRU network as IOC-MT, but had only one output head. We developed six GRU-ST models to predict FD in the six organ systems respectively. The loss function for each model was the BCE loss between predictions and labels for a single organ system.

LSTM single-task model (LSTM-ST) It was the same as GRU-ST, except that we used LSTM network instead of GRU. The LSTM is another commonly used gated variant of RNN for modeling sequential data [28].

Transformer single-task model (Transformer-ST) It was also a single-task model with one output head, while the encoder was a Transformer encoder [27] using masked multi-head attention. We provided the detail of the Transformer encoder in Appendix A.2.

Extreme gradient boosting single-task model (XGB-ST) XGB was a classic machine learning model which was composed of many basic decision trees and employed an improved boosting ensemble algorithm [29]. XGB was incapable of handling MTS data, so it only analyzed the multivariate data in the current hour (i.e. the current row vector of the concatenated matrix $X \mid\mid M$) for real-time prediction rather than analyzing the MTS from admission to the current hour.

Random forest single-task model (RF-ST) RF was another classic machine learning model which also composed of decision trees and used a bagging ensemble algorithm [30]. It also used the current-hour data for real-time prediction.

Ablation study

To evaluate the contribution of the GAT, AAM and missing indicators M to model performance, we proposed three ablation models. The first was IOC-MT without AAM, where the β value in formula (6) was fixed to be 0.5 and the computation of formula (7) was omitted. The second is IOC-MT without GAT, where the GAT module was omitted. And it should be noted that the AAM was also omitted along with GAT as the aggregated feature derived from GAT was requisite to implement the AAM (see formula (7)), so this second model was just a routine MT framework composed of a shared encoder and multiple separate output heads. The last is IOC-MT without missing indicators, where we disused the matrix of missing indicators M and just used X as model input in which the missing measurements were imputed by default zero. We compared the AUROC and AUPRC of these ablation models to the full IOC-MT in internal and external validation. In addition, we also compared their time cost per epoch during model training to evaluate the training efficiency of these models.

Experimental setup

We randomly divided the MIMIC-III dataset into the training set (80%) and the validation set (20%). The MIMIC-IV and eICU-CRD dataset were respectively used as the internal and external test set (Fig. 3). We normalized the measurements of all the datasets by removing the mean and scaling to unit variance, where the mean and variance of each variable were derived from the training set. We used the Adam optimizer [31] to iteratively tune model parameters on the training set. For each model we performed model training for 50 epochs, and saved the optimal parameters which achieved the lowest loss on the validation set. Then we evaluated model performance on the internal and external test set. For each model type, we repeated the above model training and validation experiments five times using different random initialization of model parameters to obtain the mean and 95% confidence interval (CI) of the metrics for model performance. Besides, we performed grid search to select optimal model hyperparameters that obtain better performance on the validation set, and the detail of the searching ranges for the major hyperparameters of each model was provided in Appendix A.5.

Statistical analysis and evaluation of model performance

We compared the baseline characteristics of included patients from the three databases. Continuous variables were described as mean (standard deviation) or median [interquartile range], and categorical features were described as number (percentage). The



Fig. 3 Flow chart for patient selection and experiment setup

Statistical difference was analyzed using either F-test, K-W test or Chi-square test as appropriate, and two tailed P < 0.05 was considered as statistical significance.

We evaluated both the model discrimination and calibration in the internal and external validation. The area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) were used to assess model discrimination, and the calibration curve was used to visualize model calibration [32]. To obtain the calibration curve for each model type, we plotted means of the decile-binned predicted probabilities of the five model instances versus corresponding means of actual probabilities in the patients in each bin. The calibration was assessed by inspecting the proximity between the calibration curve and the identity line of y = x which represented perfect calibration.

Results

Participants and baseline characteristics

We ultimately included 19,372 patients from MIMIC-III, 11,008 patients from MIMIC-IV, and 43,406 patients from eICU-CRD (Fig. 3), and Table 1 showed the comparison. The result showed that patients from the three databases had no statistical difference in gender, age and hospital mortality, but had difference in BMI, admission type, length of ICU stay and SOFA of the first 24 h. We also provided the overall incidence rate of FD in the six organ systems under different gap windows. In these three databases, the cardiovascular system had the highest FD rate (except the eICU-CRD under 24 h gap) and the liver system had the lowest FD rate. All organ systems showed gradually increasing FD rates when a longer gap window was set, where the cardiovascular system showed the minimal increment. The eICU-CRD showed obviously different FD rates compared to MIMIC-III and MIMIC-IV, especially that it had lower rate in CNS system and higher rate in renal system.

Table 1	Comparison of baseline characteristics
---------	--

	MIMIC-III	MIMIC-IV	elCU-CRD	P-value
	(<i>n</i> =19372)	(<i>n</i> =11008)	(n=43406)	
Gender (male), n (%)	10,926 (56.4)	6264 (56.9)	24,080 (55.5)	0.079
Age (y, mean (SD))	63.74 (16.41)	63.47 (16.08)	63.44 (16.09)	0.318
BMI (kg/m ² , mean (SD))	28.72 (7.50)	28.99 (7.82)	29.37 (8.53)	< 0.001
Admission type, n (%)				< 0.001
Medical	13,134 (67.8)	7133 (64.8)	34,440 (79.3)	
Unscheduled surgical	3797 (19.6)	3600 (32.7)	8006 (18.4)	
Scheduled surgical	2441 (12.6)	275 (2.5)	960 (2.2)	
Length of ICU stay (days, median [IQR])	3.9 [2.7, 7.1]	3.9 [2.7, 6.8]	3.0 [2.0, 5.0]	< 0.001
SOFA of first 24 h (median [IQR])	4.0 [3.0, 7.0]	5.0 [2.0, 8.0]	4.0 [2.0, 7.0]	< 0.001
Hospital mortality, n (%)	2499 (12.9)	1343 (12.2)	5447 (12.5)	0.266
Respiration deterioration rate (%)				
4 h gap	3.9	3.7	2.6	< 0.001
8 h gap	5.3	4.9	3.7	< 0.001
16 h gap	7.7	7.2	5.6	< 0.001
24 h gap	9.9	9.1	7.3	< 0.001
Coagulation deterioration rate (%)				
4 h gap	4.3	4.5	4.0	< 0.001
8 h gap	5.9	6.2	5.7	< 0.001
16 h gap	8.5	8.8	8.5	< 0.001
24 h gap	10.5	10.7	10.8	< 0.001
Liver deterioration rate (%)				
4 h gap	1.4	1.3	1.3	< 0.001
8 h gap	1.9	1.8	1.9	< 0.001
16 h gap	2.9	2.8	3.0	< 0.001
24 h gap	3.8	3.6	3.9	< 0.001
Cardiovascular deterioration rate (%)				
4 h gap	21.8	20.6	17.5	< 0.001
8 h gap	22.5	21.3	18.2	< 0.001
16 h gap	23.2	21.7	18.5	< 0.001
24 h gap	23.6	21.8	18.5	< 0.001
CNS deterioration rate (%)				
4 h gap	15.6	13.7	6.9	< 0.001
8 h gap	17.9	16.4	8.9	< 0.001
16 h gap	19.5	18.0	10.8	< 0.001
24 h gap	20.3	18.5	12.0	< 0.001
Renal deterioration rate (%)				
4 h gap	4.7	5.6	7.9	< 0.001
8 h gap	6.4	7.7	11.2	< 0.001
16 h gap	9.0	10.9	16.8	< 0.001
24 h gap	10.7	13.0	20.1	< 0.001

Model performance

The AUROC and AUPRC of all the models in external validation on eICU-CRD were summarized in Fig. 4, and their calibration curves in external validation were provided in Fig. 5. Considering limited space, the results of internal validation on MIMIC-IV were provided in Appendix A.3 and A.4. The optimal hyperparameters of our models were provided in Appendix A.5.

As Fig. 4 showed, the IOC-MT had comparable AUROC and AUPRC to LSTM-ST, GRU-ST and Transformer-ST for most organ systems under the four gap windows (even higher in some cases), and these four deep models obviously outperformed the other two machine learning models, XGB-ST and RF-ST. For different organ systems,



Fig. 4 AUROC and AUPRC for external validation

the performance of these models showed certain variation. The GRU-ST had higher AUROC and AUPRC for CNS and renal system (except under 4 h gap for renal system), but it had lower AUROC for liver system. The Transformer-ST had high AUROC for liver system, but performed worse than the other three deep models for coagulation, CNS and renal system. As the only MT model, the IOC-MT kept relatively balanced performance among the six organ systems.

Figure 5 demonstrated that all the six models had good calibration for cardiovascular system under all gap windows. For renal system, the four deep models also showed relatively good calibration, but the XGB-ST and RF-ST had biased calibration. For coagulation system, the IOC-MT, LSTM-ST and Transformer-ST showed better calibration than the other three models under 16 and 24 h gap, and under 8 and 16 h gap the IOC-MT



Fig. 5 Calibration curves for external validation

had better calibration. For respiration and CNS system, all the models showed biased calibration to a certain degree, and they overestimated the risk in most cases. At last, all the models show the poorest calibration for liver system, as they also overestimated the risk especially in the high predicted-risk bins.

In the internal validation, Appendix A.3 showed that the models had higher AUROC and AUPRC for most organ systems compared to the external validation (except the renal system). The performance difference among the models and organ systems was similar to the external validation. The calibration curves in Appendix A.4 showed that all the models had good calibration for cardiovascular system except the RF-ST, and these models also had good calibration for renal system except the Transformer-ST and RF-ST. Besides, all the models except RF-ST had better calibration for CNS system in the internal validation than in the external test, and the four deep learning models also had better calibration for liver system.

Ablation study

Tables 2 and 3 compared the AUROC and AUPRC of the ablation models to the full IOC-MT in external and internal validation under 4 h gap, and their time cost per epoch for model training were provided in the rightmost column of the tables. The results under 8, 16, 24 h gap were provided in Appendix A.6. Our results showed that in most cases, the full IOC-MT performed better than the ablation model without AAM, especially for the CNS and renal system, and the ablation model without AAM performed better than the model without GAT. This showed the contribution of GAT and AAM to improving the model's overall performance for MT prediction. Besides, the full IOC-MT also outperformed the ablation model without missing indicators, which indicated the importance of marking data missingness. The result of training time showed that introducing GAT in routine MT framework prolonged the training time by about 8–10 s per epoch under different gap windows (Without AAM vs. Without GAT), and further introducing AAM prolonged the time by about 3–5 s per epoch (Full model vs. Without AAM).

Adaptive adjustment mechanism in IOC-MT

Figure 6 showed the overall α and β values of an IOC-MT model in the external validation under 4 h gap. For all the six organ systems, the sample proportion of the 0.0-0.2 β group was the largest. This indicated that in most cases the IOC-MT only made slight adjustment to the original feature. In the 0.4–0.6, 0.6–0.8 and 0.8-1.0 β group where the model made more adjustment, the six organ systems had different attention assignments. Specifically, respiration system mainly relied on CNS system; coagulation system relied more on liver system; liver system mainly relied on respiration, coagulation and CNS system; cardiovascular system mainly relied on liver system; CNS system relied on all the other organ systems in 0.4–0.6 and 0.6–0.8 β group, and mainly relied on liver system in 0.8-1.0 β group; renal system relied more on coagulation and cardiovascular system. It should be noted that in Fig. 6 each organ system assigned zero attention weight on itself as the original feature of itself was not included in its own information aggregation (Formula (4)).

At individual level, we selected a patient staying in ICU for 55 h from the external test set, and used the same IOC-MT model to perform dynamic prediction for the six organ

Table 2 🗚	Ablation study fo	r external va	lidation under	.4h gap								
	Respiration		Coagulation		Liver		Cardiovas	cular	CNS		Renal	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC

	Respiration		Coagulation		Liver		Cardiovascu	ar	CNS		Renal		Training time (s/
	AUROC	AUPRC	epoch)										
Without AAM	0.722 ± 0.007	0.080 ± 0.001	0.813 ± 0.009	0.145 ± 0.001	0.777 ± 0.007	0.053 ± 0.001	0.826 ± 0.001	0.510 ± 0.002	0.646 ± 0.001	0.131 ± 0.001	0.737±0.016	0.185 ± 0.007	27.4±0.1
Without GAT	0.720 ± 0.002	0.073 ± 0.005	0.804 ± 0.001	0.143 ± 0.003	0.762 ± 0.002	0.050 ± 0.003	0.821 ± 0.004	0.489 ± 0.010	0.634 ± 0.004	0.124 ± 0.004	0.731 ± 0.013	0.186 ± 0.006	18.4±0.1
Without missing indicators	0.723±0.006	0.081 ± 0.002	0.777 ± 0.005	0.125 ± 0.002	0.761 ± 0.005	0.049±0.002	0.810±0.001	0.483 ± 0.003	0.632 ± 0.004	0.125 ± 0.001	0.800 ± 0.009	0.249±0.019	34.8±0.6
Full model	0.729±0.005	0.085±0.001	0.813±0.003	0.139±0.006	0.774 ± 0.009	0.055±0.003	0.829±0.003	0.517±0.008	0.687±0.004	0.147±0.004	0.851±0.016	0.348±0.019	32.5 ± 0.4

systems under 4 h gap. The result was showed in Fig. 7. The label at each hour could be inferred by the SOFA plot. For instance, the SOFA for respiration system was 0 at the 1st hour and the maximum SOFA in corresponding prediction window (6th to 9th hour) was 3, so the label at the 1st hour was deteriorating (positive). As the green square in the subplot of respiration system showed, the IOC-MT output up-adjusted predicted risks compared to its original prediction from 2nd to 4th hour, and the β values in these hours were also higher than other hours. The lower-half subplot showed that the adjustment mainly relied on liver and renal system. This adjustment made the model correctly predict the coming FD. Similar up adjustment was seen in the green squares in the coagulation and renal subplot. The most significant adjustment occurred in cardiovascular system. In the two green squares, the SOFA score fluctuated between 0 and 1 (or 3 at 11th hour), where most 0-score hours had positive label and 1- and 3-score hours had negative label. The original predicted risk kept high at all these hours, while the IOC-MT adjusted the risk down to almost zero at the 1- and 3-score hours. The β values at these hours were more than 0.6 and the liver, renal and coagulation system contributed most to these adjustments. At last, for liver and CNS, most β values along the time series were very low and there was almost no difference between the original and adjusted predicted risks.

Discussion

In this study we proposed a multi-task deep learning framework named IOC-MT for dynamically predicting FD in multiple organ systems. The IOC-MT took in MTS data of routinely monitored clinical variables in ICU, and output hourly predictions for patient's six major organ systems simultaneously. The major contribution of this study was that we used the GAT module to model inter-organ correlation and introduced the AAM to adjust prediction. Our experimental results showed that the IOC-MT had comparable performance to the classic ST deep learning models and outperformed the routine MT framework using separate output heads. In addition, the AAM of our IOC-MT was comprehensible through its intuitive attention weights α and adjustment coefficient β , rather than be like a black box.

This study used the SOFA score to define organ dysfunction because: (1) it is the most commonly used scoring criterion for assessing the function of the most critical six organs, especially that sepsis 3.0 use it to define organ dysfunction [33]; (2) compared with other similar criteria, such as the Multiple Organ Dysfunction (MOD) score [34] and the Logistic Organ Dysfunction (LOD) score [35], SOFA shows similar or better prognostic value [36-38], and meanwhile it is more applicable as all its component variables are easily accessible; (3) it is convenient for clinicians to track the changes of organ function by regularly repeated scoring of SOFA. Generally, the SOFA score is calculated every 24-48 h in ICU [39], but we calculate hourly SOFA in this study in order to improve the timeliness of our model, since more frequent prediction enable clinicians to receive warning and start intervention timely. In a previous related study, cardiovascular dysfunction is defined as the onset of vasoactive medication and respiratory dysfunction is defined as the onset of mechanical ventilation [17]. Compared to SOFA score, such definitions take insufficient consideration of the early signs for organ dysfunction, such as fall of blood pressure or oxygenation index, so their trained model is unable to recognize early FD.

	Renal
	CNS
	Cardiovascular
	Liver
ial validation under 4 h gap	Coagulation
e 3 Ablation study for intern	Respiration
Tab	

	Respiration		Coagulatior		Liver		Cardiovascu	ılar	CNS		Renal		Training time (s/
	AUROC	AUPRC	epoch)										
Without AAM	0.762 ± 0.003	0.151 ± 0.005	0.838 ± 0.004	0.194 ± 0.004	0.831 ± 0.004	0.064 ± 0.001	0.818 ± 0.001	0.545 ± 0.003	0.699 ± 0.007	0.324 ± 0.011	0.788 ± 0.017	0.196 ± 0.021	27.4±0.1
Without GAT	0.756 ± 0.009	0.149 ± 0.003	0.832 ± 0.003	0.176 ± 0.007	0.827 ± 0.007	0.065 ± 0.005	0.813 ± 0.001	0.530 ± 0.002	0.673 ± 0.008	0.299 ± 0.006	0.786 ± 0.006	0.192 ± 0.012	18.4±0.1
Without missing indicators	0.754±0.013	0.148±0.003	0.813±0.002	0.175 ± 0.003	0.820 ± 0.001	0.062 ± 0.002	0.816±0.002	0.540 ± 0.005	0.662 ± 0.002	0.275 ± 0.001	0.797 ± 0.003	0.220±0.007	34.8±0.6
Full model	0.786 ± 0.006	0.159 ± 0.003	0.838±0.001	0.181 ± 0.005	0.835±0.006	0.067±0.002	0.835±0.001	0.579±0.001	0.731 ± 0.005	0.358±0.007	0.835±0.016	0.300 ± 0.014	32.5 ± 0.4

In this study we compare our IOC-MT to many baseline models, including three ST sequential deep models (LSTM, GRU, Transformer) and two ST non-sequential machine learning models (XGB, RF). In addition, we also assess three MT ablation models of the IOC-MT. All the models are validated on the internal and external test set, which increases the persuasiveness of our study. These models perform better in internal validation than in external validation. This is mainly because that the data difference between eICU-CRD and MIMIC-III is greater than that between MIMIC-IV and MIMIC-III (Table 1). The sequential models perform much better than the nonsequential models, especially in the external validation, which shows their advantage for handling clinical MTS data and excellent generalizability for multi-center application. We prove that the GAT and AAM improve the overall performance of the routine multiple output-head MT framework, despite that they increase the time for model training time. The missing indicators are effective to handle data missingness in this study, as it not only improves the model performance but also slightly reduce the training time. We think that our proposed model architecture and related methods are also applicable to other scene of clinical MT prediction.

Although the IOC-MT achieve comparable performance to the ST deep models, its AUROC and AUPRC are still slightly lower in most cases of our experiments, and such a performance degradation is more significant for the ablation model without GAT (i.e. routine MT framework). This issue is referred to as negative transfer which is common for MT deep models [14]. The intrinsic reason for negative transfer is that training MT model is not as flexible as training ST model since the MT model needs to use one encoder to learn the shared feature for all the tasks. To mitigate negative transfer, a sort of method is to assign different subsets of parameters in the shared encoder to each task, which enable the shared encoder to encode multiple task-specific features rather than one shared feature. The previous SeqSNR [17] is belong to this method. We argue that this method may be more appropriate when the correlation of multiple tasks is not so close since it essentially use relatively independent components of the whole MT model to handle different tasks. Unlike this method, our method focuses on the output heads and their correlation, which is based on the professional knowledge that the organ systems are correlated. Our ablation study proves that introducing GAT and AAM can mitigate negative transfer of the routine MT framework.

The IOC-MT captures inter-organ correlations by the attention weights α and adjustment coefficient β . From the clinical perspective, these data-based correlations should have biological plausibility. When an organ relies on another for predicting itself, there should be reasonable causal relationship between them. Our results prove that in most cases IOC-MT has the biological plausibility. For instance, Fig. 6 shows that the respiration system mainly relies on CNS system in high β groups. This is reasonable as many CNS diseases can lead to respiratory dysfunction [40], so the information of CNS system should be valuable for predicting FD in respiration system. Similarly, coagulation system mainly relies on liver system (Fig. 6), which is also reasonable as liver is the major organ to synthesize coagulation factors and liver diseases often affect coagulation system [41]. Figure 6 also shows that the model predicts renal function based on cardiovascular system. As we know, circulation failure and hypotension can cause hypoperfusion of kidney and lead to AKI [42]. Then the renal and cardiovascular subplots of Fig. 7 further prove this case. From 38th to 42nd hour, the IOC-MT adjusts the FD risk for renal system



Fig. 6 The α and β values of an IOC-MT in the external validation under 4 h gap. Each subplot was for an organ system. The left half of each subplot was the sample proportion in each β group, and the right half was the average attention weights in each β group. For intuitive presentation, we used the abbreviation of organ to replace the mathematical symbol of attention weight. For instance, 'Coag-> Resp' represented the attention weight α_{21} which indicated the importance of coagulation system for predicting respiration function

upwards, and the adjustment is mainly based on cardiovascular system. Meanwhile, the cardiovascular SOFA of this patient is 1 during this period, indicating that he suffers from cardiovascular dysfunction. Thus, the IOC-MT successfully captures the correlation that cardiovascular system can affect renal system. However, it should be noted that not all the correlations are so reasonable. For instance, Fig. 6 shows that the cardiovascular system relies less on respiration system but more on liver system as β increases. Nevertheless, as we know, the biological correlation between the respiration and cardiovascular system should be closer.

Our study has several limitations. Firstly, although data missingness is inevitable in our study, the method that uses previous SOFA score to fill the current missingness may produce unreliable labels, especially when the related variable is missing for a long period. For instance, the FD rate of liver system is very low in Table 1, but we find that the total bilirubin is relatively less recorded in the three databases (even only once in 1–2 weeks). If there is elevated bilirubin which is not recorded (this possibility will increase as the missing period prolongs), the actual liver FD rate is underestimated and many real-time labels defined by our method may be false negative. This will cause biased prediction of our model. Collecting more high-quality data is feasible for solving this issue. Secondly, we have not made modification on the shared encoder of our MT framework as this study focuses on modeling inter-organ correlation for MT prediction. We think that it is also promising to propose reasonable improvement for the shared encoder and we will perform further research in this direction. At last, our results show that the IOC-MT needs additional training time compared to routine MT-framework. Although the increment of time cost is still acceptable as there were only six nodes in our GAT, it



Fig. 7 Dynamic predictions for six organ systems of a patient by an IOC-MT under 4 h gap. Each subplot was for an organ system. The upper half of each subplot provided the hourly original and adjusted predicted risks, as well as hourly β values and SOFA score for this organ. The lower half was the attention weights for information aggregation of this organ at each hour. The upper and lower half were aligned along the time axis, and the green square marked out the difference between the original and adjusted prediction

will be computational expensive as tasks increase. Some previous studies propose efficient modification on attention algorithm of GAT [43, 44], and we will perform related research to improve the efficiency of IOC-MT in our future work.

Conclusion

The IOC-MT is a promising deep learning framework for predicting FD in six organ systems. It can capture inter-organ correlation and adjust the prediction based on interorgan information aggregation. The IOC-MT has comparable performance to ST deep models and outperforms routine MT framework. The AAM of IOC-MT is intuitive and comprehensible.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-025-00445-w.

Supplementary Material 1

Acknowledgements

None.

Author contributions

XG conceived the idea and the study design, performed algorithm program, and revised the manuscript. ZXZ and YL performed literature review, data collection and manuscript writing. SY, XC, WBN and YYX helped to collect data. MJL helped to revise English writing. All authors read and approved the final manuscript.

Funding

This study was supported by Natural Science Foundation of Hunan Province of China (2023JJ60079, Xun Gong; 2022JJ30796, Minjie Lin), National Natural Science Foundation of China (82100624, Minjie Lin), Beijing Union Medical Fund - Rui E (Ruiyi) Emergency Medical Research Fund (22222012012, Xun Gong), and Scientific Research Program of Hunan Provincial Health Commission (D202310007453, Xun Gong). Minjie Lin was supported by the Central South University Postdoctoral Programme.

Availability of data and codes

Data of MIMIC-III is available on website at https://physionet.org/content/mimiciii-demo/1.4/; data of MIMIC-IV is available on website at https://physionet.org/content/mimiciv/1.0/; data of eICU-CRD is available on website at https://eicu-crd.mit.edu/. The codes for data preparation and model implementation are available at https://github.com/gongxu n1246/IOC-MT.

Declarations

Ethics approval and consent to participate

This study was an analysis of third-party deidentified publicly available databases with pre-existing ethical review board (ERB) approval.

Competing interests

The authors declare no competing interests.

Received: 22 October 2024 / Accepted: 4 April 2025

Published online: 16 April 2025

References

- Ferreira AM, Sakr Y. Organ dysfunction: general approach, epidemiology, and organ failure scores. Semin Respir Crit Care Med. 2011;32(5):543–51. https://doi.org/10.1055/s-0031-1287862.
- Sakr Y, Lobo SM, Moreno RP, et al. Patterns and early evolution of organ failure in the intensive care unit and their relation to outcome. Crit Care. 2012;16(6):R222. https://doi.org/10.1186/cc11868. Published 2012 Nov 16.
- Beal AL, Cerra FB. Multiple organ failure syndrome in the 1990s. Systemic inflammatory response and organ dysfunction. JAMA. 1994;271(3):226–33.
- Schuler A, Wulf DA, Lu Y, et al. The impact of acute organ dysfunction on Long-Term survival in sepsis. Crit Care Med. 2018;46(6):843–9. https://doi.org/10.1097/CCM.00000000003023.
- Shapiro N, Howell MD, Bates DW, Angus DC, Ngo L, Talmor D. The association of sepsis syndrome and organ dysfunction with mortality in emergency department patients with suspected infection. Ann Emerg Med. 2006;48(5):583–e5901. https ://doi.org/10.1016/j.annemergmed.2006.07.007.
- Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116–9. https://doi.org/10.1038/s41586-019-1390-1.

- Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med. 2020;26(3):364–73. https://doi.org/10.1038/s41591-020-0789-4.
- Bolourani S, Brenner M, Wang P, et al. A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation. J Med Internet Res. 2021;23(2):e24246. https://doi.org/10.21 96/24246. Published 2021 Feb 10.
- Liu C, Yao Z, Liu P, et al. Early prediction of MODS interventions in the intensive care unit using machine learning. J Big Data. 2023;10(1):55. https://doi.org/10.1186/s40537-023-00719-2.
- 10. Liu G, Xu J, Wang C, et al. A machine learning method for predicting the probability of MODS using only non-invasive parameters. Comput Methods Programs Biomed. 2022;227:107236. https://doi.org/10.1016/j.cmpb.2022.107236.
- 11. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med. 2018;6(12):905–14. https://doi.org/10.1016/S2213-2600(18)30300-X.
- 12. Xue B, Li D, Lu C, King CR, Wildes T, Avidan MS, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. JAMA Netw Open. 2021;4(3):e212240. ht tps://doi.org/10.1001/jamanetworkopen.2021.2240
- 13. Ruder S. An overview of Multi-Task learning in deep neural networks. 2017. https://doi.org/10.48550/arXiv.1706.05098
- Vandenhende S, Georgoulis S, Gansbeke WV, et al. Multi-Task learning for dense prediction tasks: A survey. IEEE Trans Pattern Anal Mach Intell. 2021;01https://doi.org/10.1109/TPAMI.2021.3054719.
- Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data. 2019;6(1):96. https://doi.org/10.1038/s41597-019-0103-9. Published 2019 Jun 17.
- Cherifa M, Interian Y, Blet A, Resche-Rigon M, Pirracchio R. The physiological deep learner: first application of multitask deep learning to predict hypotension in critically ill patients. Artif Intell Med. 2021;118:102118. https://doi.org/10.1016/j.ar tmed.2021.102118.
- 17. Roy S, Mincu D, Loreaux E, et al. Multitask prediction of organ dysfunction in the intensive care unit using sequential subnetwork routing. J Am Med Inf Assoc. 2021;28(9):1936–46. https://doi.org/10.1093/jamia/ocab101.
- Panwar R, Tarvade S, Lanyon N, et al. Relative hypotension and adverse Kidney-related outcomes among critically ill
 patients with shock. A multicenter, prospective cohort study. Am J Respir Crit Care Med. 2020;202(10):1407–18. https://doi
 .org/10.1164/rccm.201912-2316OC.
- Darmon M, Clec'h C, Adrie C, et al. Acute respiratory distress syndrome and risk of AKI among critically ill patients. Clin J Am Soc Nephrol. 2014;9(8):1347–53. https://doi.org/10.2215/CJN.08300813.
- Matsuura R, Doi K, Rabb H. Acute kidney injury and distant organ dysfunction-network system analysis. Kidney Int. 2023;103(6):1041–55. https://doi.org/10.1016/j.kint.2023.03.025.
- 21. Elikovi P, Cucurull G, Casanova A, et al. Graph Atten Networks. 2017. https://doi.org/10.48550/arXiv.1710.10903.
- 22. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.
- Goldberger AL, Amaral LA, Glas L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):E215–20. https://doi.org/10. 1161/01.cir.101.23.e215.
- 24. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The elCU collaborative research database, a freely available multi-center database for critical care research. Sci Data. 2018;5(1):180178. https://doi.org/10.1038/sdata.2018.178.
- Vincent JL, de Mendonça A, Cantraine F, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Working group on sepsis-related problems of the European society of intensive care medicine. Crit Care Med. 1998;26(11):1793–800. https://doi.org/10.1097/00003246-1998110 00-00016.
- Moreno R, Vincent JL, Matos R, et al. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. Working group on sepsis related problems of the ESICM. Intensive Care Med. 1999;25(7):686–96. https://doi.org/10.1007/s001340050931.
- 27. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Adv Neural Inf Process Syst 2017; 30.
- Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco. 1997;98,1735.
- 29. Chen T, Guestrin C, Xgboost. A scalable tree boosting system. ArXiv. New York: ACM; 2016. p.785-94.
- 30. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
- 31. Kingma DP, Ba J, Adam. A Method for Stochastic Optimization. In International Conference on Learning Representations, 2015.
- 32. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA. 2017;318(14):1377–84. https://doi.org/10.1001/jama.2017.12126.
- Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016;315(8):801–10. https://doi.org/10.1001/jama.2016.0287.
- Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Crit Care Med. 1995;23(10):1638–52. https://doi.org/10.1097/00003246-199510000-00007.
- 35. Le Gall JR, Klar J, Lemeshow S, et al. The logistic organ dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU scoring group. JAMA. 1996;276(10):802–10. https://doi.org/10.1001/jama.276.10.802.
- 36. Zygun D, Berthiaume L, Laupland K, Kortbeek J, Doig C. SOFA is superior to MOD score for the determination of nonneurologic organ dysfunction in patients with severe traumatic brain injury: a cohort study. Crit Care. 2006;10(4):R115. https://doi.org/10.1186/cc5007.
- Peres Bota D, Melot C, Lopes Ferreira F, Nguyen Ba V, Vincent JL. The multiple organ dysfunction score (MODS) versus the sequential organ failure assessment (SOFA) score in outcome prediction. Intensive Care Med. 2002;28(11):1619–24. https://doi.org/10.1007/s00134-002-1491-3.
- Timsit JF, Fosse JP, Troché G, et al. Calibration and discrimination by daily logistic organ dysfunction scoring comparatively with daily sequential organ failure assessment scoring for predicting hospital mortality in critically ill patients. Crit Care Med. 2002;30(9):2003–13. https://doi.org/10.1097/00003246-200209000-00009.

- Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related organ failure Assessment) score to describe organ dysfunction/failure. On behalf of the working group on Sepsis-Related problems of the European society of intensive care medicine. Intensive Care Med. 1996;22(7):707–10. https://doi.org/10.1007/BF01709751.
- 40. Li X, Deng J, Long Y, et al. Focus on brain-lung crosstalk: preventing or treating the pathological vicious circle between the brain and the lung. Neurochem Int. 2024;178:105768. https://doi.org/10.1016/j.neuint.2024.105768.
- 41. Franchini M, Mannucci PM. Coagulation abnormalities in chronic liver disease. Semin thromb Hemost. Published Online Febr. 2025;27. https://doi.org/10.1055/a-2531-4712.
- 42. Saugel B, Sander M, Katzer C, et al. Association of intraoperative hypotension and cumulative norepinephrine dose with postoperative acute kidney injury in patients having noncardiac surgery: a retrospective cohort analysis. Br J Anaesth. 2025;134(1):54–62. https://doi.org/10.1016/j.bja.2024.11.005.
- Fofanah AJ, Chen D, Wen L, et al. Addressing imbalance in graph datasets: introducing GATE-GNN with graph ensemble weight attention and transfer learning for enhanced node classification. Expert Syst Appl. 2024;255. https://doi.org/10.101 6/j.eswa.2024.124602.
- 44. Fofanah AJ, Leigh AO. EATSA-GNN: Edge-Aware and Two-Stage attention for enhancing graph neural networks based on teacher–student mechanisms for graph node classification. Neurocomputing. 2025;612. https://doi.org/10.1016/j.neucom .2024.128686.

Publisher's note

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