DOI: 10.16781/j.CN31-2187/R.20240183

·论　著·

基于循环神经网络模型的创伤重症患者临床结局的动态预测

齐戈尧[1△]，徐　进[2△]，金志超[1*]
1. 海军军医大学（第二军医大学）卫生勤务学系军队卫生与统计学教研室，上海 200433
2. 中国人民解放军联勤保障部队第九四〇医院，兰州 730030

［摘要］ **目的**　探讨基于循环神经网络（RNN）算法构建的动态预测模型用于创伤重症患者临床结局动态预测的价值，并研究动态策略和实时预测模型可行的搭建方案及路径。**方法**　本研究数据来源于美国重症监护医学信息数据库（MIMIC）-Ⅳ 2.0。以创伤重症患者院内结局为预测目标，使用长短时记忆（LSTM）和门控循环单元（GRU）2种RNN算法分别在4、6和8 h时间窗下训练动态预测模型。使用灵敏度、特异度、F1值和AUC值对模型性能进行评价，并分析不同RNN算法和时间窗对模型性能的影响。在8 h时间窗下分别训练隐马尔科夫模型（HMM）、随机森林（RF）模型和logistic模型作为对照，横向比较2种RNN算法模型与对照模型的性能指标，并分析各模型的时间趋势变化。**结果**　在不同时间窗时，RNN动态模型在灵敏度、特异度、F1值和AUC值等4个性能指标上差异均有统计学意义（均$P<0.001$），在8 h时间窗时模型的各性能指标均高于6 h和4 h时；不同RNN算法（LSTM和GRU）间仅特异度差异有统计学意义（$P=0.036$）。横向比较结果显示，2种RNN算法模型和其他模型间各性能指标差异均有统计学意义（均$P<0.001$），2种RNN算法模型各指标均高于HMM、RF和logistic模型；各算法模型灵敏度、特异度和F1值的ICC均小于0.400（95% *CI*未包含0），而AUC值的ICC在统计学上证据不足（95% *CI*包含0）。**结论**　基于RNN算法的动态模型对创伤重症患者临床结局的预测效果较其他常见模型具有一定优势，且时间窗对模型性能可能存在影响。

［关键词］ 循环神经网络；长短期记忆网络；门控循环单元；创伤；动态模型；临床结局；预测模型
［引用本文］ 齐戈尧，徐进，金志超. 基于循环神经网络模型的创伤重症患者临床结局的动态预测[J]. 海军军医大学学报，2024，45（10）：1241-1249. DOI: 10.16781/j.CN31-2187/R.20240183.

Dynamic prediction of clinical outcomes for critical trauma patients based on a recurrent neural network model

QI Geyao[1△], XU Jin[2△], JIN Zhichao[1*]
1. Department of Military Health Statistics, Faculty of Health Services, Naval Medical University (Second Military Medical University), Shanghai 200433, China
2. No. 940 Hospital of Joint Logistics Support Force of PLA, Lanzhou 730030, Gansu, China

［**Abstract**］ **Objective**　To explore the value of dynamic prediction model based on recurrent neural network (RNN) algorithms for dynamic prediction of clinical outcomes in patients with critical trauma, and to study the feasible construction scheme and path of dynamic strategy and real-time prediction model. **Methods**　The data of this study were derived from the US Medical Information Mart for Intensive Care (MIMIC) - Ⅳ 2.0. In order to predict the in-hospital outcomes of critical trauma patients, 2 RNN algorithms, long short-term memory (LSTM) and gated recurrent unit (GRU) were used to train dynamic prediction models under the time windows of 4, 6 and 8 h, respectively. The performance of the models was evaluated using the sensitivity, specificity, F1 value and area under curve (AUC) value; and the effects of different RNN algorithms and time windows on the performance of the models were analyzed. Hidden Markov model (HMM), random forest (RF) model and logistic model were trained under 8-h time window as the controls to compare the performances and the time trends horizontally with the 2 RNN algorithm models. **Results**　There were significant differences in the 4 performance indexes of the RNN dynamic models including the sensitivity, specificity, F1 value and AUC value (all $P<0.001$), and the performance indexes at 8-h time window were higher than those at 6 h and 4 h; there was only significant difference in specificity between different RNN algorithms (LSTM & GRU) ($P=0.036$). The results of the horizontal comparison showed

［收稿日期］ 2024-03-22　［接受日期］ 2024-08-26
［基金项目］ 上海市卫生健康委员会卫生行业临床研究专项（202340037），海军军医大学（第二军医大学）“三航”计划. Supported by Clinical Research Project of Health Industry of Shanghai Municipal Health Commission (202340037) and “San Hang” Program of Naval Medical University (Second Military Medical University).
［作者简介］ 齐戈尧，硕士，主治医师. E-mail: adsuver@sina.com；徐　进，副主任医师. E-mail: 419339187@qq.com
△共同第一作者（Co-first authors）.
*通信作者（Corresponding author）. Tel: 021-81871442, E-mail: jinzhichao@smmu.edu.cn

that there were significant differences in each performance index between the 2 RNN prediction models and other models (all $P<0.001$), and each index of the 2 RNN algorithm models was higher than those of the HMM, RF model and logistic model. The intraclass correlation coefficients (ICCs) of each algorithmic model were less than 0.400 for the sensitivity, specificity and F1 value (0 was not included in 95% confidence interval［*CI*］), while the ICCs for the AUC value were statistically under-evidenced (0 was included in 95% *CI*). **Conclusion**　The dynamic models based on RNN algorithms have certain performance advantages over those based on other common algorithms, and the time window may have an impact on the model performance.

［**Key words**］ recurrent neural network; long short-term memory; gated recurrent unit; trauma; dynamic model; clinical outcomes; predicting model

［**Citation**］ QI G, XU J, JIN Z. Dynamic prediction of clinical outcomes for critical trauma patients based on a recurrent neural network model[J]. Acad J Naval Med Univ, 2024, 45(10): 1241-1249. DOI: 10.16781/j.CN31-2187/R.20240183.

创伤重症是指患者受伤后并由此引发的一系列危及生命的急重综合征，包括失血性休克、开放性气胸、脏器破裂等[1]。据报道，2020 年我国因道路交通事故导致的死亡人数达 6 万余人[2]，美国每年有超过 6 万例患者死于创伤失血性休克，而全球范围则超过了 150 万例[3]，可以说创伤及其引起的重症并发症已经成为意外死亡的重要原因之一。创伤重症具有转归复杂、进展急骤等特点，因此在对创伤重症患者救治的临床实践中，救治团队的处置经验和临床判断对抢救成功与否尤为重要[4]。充分利用创伤重症抢救的数据，开发一个可以对创伤重症患者病程进展和临床结局进行实时预测的动态预测模型用以辅助医疗决策，对挽救伤员生命具有重要意义。

循环神经网络（recurrent neural network，RNN）算法是目前较为成熟的动态预测模型之一。RNN 在各领域有着广泛的应用，如自然语言处理、股价波动等动态时序预测[5]，其模型性能也充分接受了行业实践的检验[6]。本研究以 RNN 及其衍生算法为基础，构建基于真实创伤重症病例的动态预测模型，对患者的临床结局进行实时预测，分析探讨基于 RNN 算法的动态预测模型在创伤重症患者临床辅助决策中的应用价值。

1　资料和方法

1.1　数据资料与病例筛选

1.1.1　数据来源　美国重症监护医学信息数据库（Medical Information Mart for Intensive Care，MIMIC）成立于 2003 年，是一个受美国国立卫生研究院资助的单中心、长周期、大样本且免费开放的公共数据库[7]。目前第四版数据库 MIMIC-Ⅳ中共收录了超过 19 万例患者的临床数据及超过 45 万例次的住院信息。本研究数据来源于 MIMIC-Ⅳ 2.0。本研究已获得该数据库的使用权限（用于非商业用途的科学研究）。

1.1.2　研究资料　（1）入选伤情：以国际疾病分类（ICD-10）编码筛选 MIMIC-Ⅳ数据库中入院诊断为交通事故伤、撞击伤、锐器伤、火器伤和烧伤等且进入 ICU 的患者（入院时间 2010－2019 年）。

（2）采集特征：包括性别、人种、年龄、身高、体重、心率、呼吸频率、体温、收缩压、舒张压、血红蛋白、白细胞计数、血小板计数、丙氨酸转氨酶、天冬氨酸转氨酶、总胆红素、血肌酐、血尿素氮、氧分压、二氧化碳分压和血氧饱和度等共 21 个变量。

（3）结局变量：临床结局（死亡与否）、结局发生时间。

（4）纳入标准：年龄＞16 岁；筛选数据范围为患者进入 ICU 前 1 d 至结局事件出现或进入 ICU 后 7 d 内。

（5）排除标准：结局变量记录不全；总缺失数据高于 25%；单一基线变量数据缺失高于 10%。

（6）缺失处理：循环节点前后 6 h 内有数据反馈采用就近填补方式；否则采用线性均值填补方式。

1.2　预测策略与评价标准

1.2.1　特征重要性分析　在训练 RNN 动态预测模型之前，需考虑纳入模型训练的分析变量选取问题。为提升动态模型预测效能，尽可能降低无关变量对模型造成噪声干扰，以最大限度地保留变量在模型训练中的真实贡献，同时也为降低模型训练难度、节约时间成本，本研究使用 3 种不同的机器学习方法对基线变量的特征重要性指数进行计算。首

先，每种机器学习方法计算 5 次特征重要性指数，并以这 5 次的算术平均数作为该算法下的最终特征重要性指数；其次，计算 3 种机器学习算法得到的平均特征重要性指数的组内相关系数（intraclass correlation coefficient，ICC），用以评估不同算法下特征重要性的一致性，并将不同算法获得的特征重要性指数再次进行平均，最终得到汇总的特征重要性指数；最后，将各变量汇总的特征重要性指数进行排序，并以此顺序作为分析变量选择的重要依据。

考虑到分析变量间可能存在的内生性和共线性问题，在对变量进行特征重要性排序后，通过咨询临床专家（5 名来自不同三甲医院具有副高及以上职称的急诊或重症医学科临床医师）并结合过往类似研究中的有关结论和策略，从特征重要性指数较高且共线性较低的基线变量中选择若干适宜变量，作为分析变量用于动态预测模型的训练。

1.2.2　基于 RNN 算法的动态预测模型搭建　RNN 算法及其衍生算法由若干基本循环单元组成。其中，RNN 的基本循环单元主要包括输入层（x）、隐藏层（s）、输出层（y）、激活函数（f、g）和权重矩阵（$\boldsymbol{U}$、$\boldsymbol{V}$、$\boldsymbol{W}$）等组件[8]。各循环单元按时间顺序展开，通过纳入不同时间节点的分析数据，达到使用时间序列数据进行动态预测的效果。然而，基础 RNN 算法由于基本循环单元内部设计简单，在预测复杂、长时程的事件时会出现“长程依赖”问题[9-10]。为了克服上述问题，通过精细循环单元结构、增加必要的单元组件等方式形成了以长短时记忆（long short-term memory，LSTM）算法和门控循环单元（gated recurrent unit，GRU）算法为代表的 RNN 衍生算法。本研究分别使用 LSTM 算法和 GRU 算法搭建动态预测模型[11]。为强化 RNN 动态预测模型在每个循环节点间的时间关联性，本研究在各循环节点建模并在预测结束后将模型参数传递到下一节点，使得下一节点的模型在训练时都会以上一节点参数为“蓝本”继续优化。这种参数“继承”的建模方案将有利于提升 RNN 算法动态模型的预测效能，节约模型训练时间[12]。

本研究以进入 ICU 后 4 h 为起始循环节点，在每个循环节点均训练 1 个 RNN（LSTM 或 GRU）动态预测模型，训练完毕后将参数传递至下一节点，节点间隔 2 h，最长使用进入 ICU 后 168 h 的临床数据。在每个循环节点用以训练模型的时间序列数据跨度分别为 4、6、8 h（即时间窗为 4、6、8 h）。本研究将前文所述筛选纳入的病例样本通过 9 ∶ 1 的比例随机划分为训练集和测试集，其中训练集数据用于在每个循环节点训练模型（建模过程使用 5 折交叉验证策略，即各模型均需要完整迭代 5 次），每次训练时均须将队列脱落样本从训练集剔除。

1.2.3　模型评价指标和模型评价　测试集数据用于模型评价。模型在每个循环节点的预测目标均为患者的院内临床结局。模型在每个循环节点预测完毕后，针对患者临床结局预测的正误情况，分别计算该节点下预测模型的灵敏度、特异度、F1 值和 ROC AUC 值[13-14]等性能评价指标。最后，连续记录各个节点下预测模型的 4 个性能评价指标，并分析各指标的总体差异和时间趋势一致性。本研究的试验路径如图 1 所示。

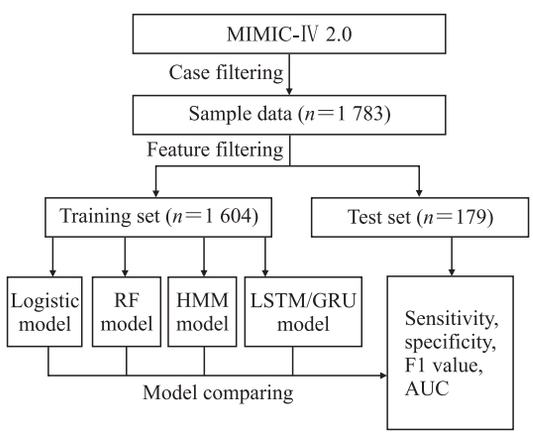

图 1　RNN 算法模型研究路径示意图

Fig 1　Experimental path of RNN algorithm model

RNN: Recurrent neural network; MIMIC: Medical Information Mart for Intensive Care; RF: Random forest; HMM: Hidden Markov model; LSTM: Long short-term memory; GRU: Gated recurrent unit; AUC: Area under curve.

本研究还将分别训练隐马尔科夫模型（hidden Markov model，HMM）[15]、随机森林模型（random forest，RF）[16]和 logistic 模型 3 种不同算法模型作为对照，用以分析 RNN 算法模型与其他预测模型的性能差异。其中，HMM 模型同样为动态模型，训练策略和训练数据同 LSTM 和 GRU 算法模型；RF 和 logistic 模型为常用的静态模型，两者仅训练 1 次，训练数据为进入 ICU 后的首次在院数据。

1.3　统计学处理　所有深度学习算法和数据统计分析均使用 Python 3.6.0 软件的 PyTorch、scikit-learn、HMMlearn 和 statsmodels 等程序包完成。计算变量特征重要性指数的方法为极端梯度提升算法（extreme gradient boosting，XGBoost）、自适应提升算法（adaptive boosting，AdaBoost）和 RF 等 3 种机器学习算法[17-18]。计数资料以例数和百分数表示，组间比较采用 χ^2 检验；计量资料以 $\bar{x}\pm s$ 表示，两组间比较采用独立样本 t 检验，多组间差异性检验使用单因素方差分析。灵敏度、特异度、F1 值和 AUC 值等指标的模型间评价使用析因设计的方差分析（包括主效应和交互效应）。纳入变量的特征重要性和模型评价指标在不同算法模型间的一致性评价指标采用 ICC（计算模式包括双向混合、一致性、单一度量）[19]。所有检验均为双侧检验，检验水准（α）为 0.05。

2　结　果

2.1　创伤重症患者基本情况　按照病例筛选标准，共 1 783 例创伤重症患者入组，死亡病例 262 例（14.69%），未死亡病例 1 521 例（85.31%），平均入住 ICU 时间为（162.096±55.268）h，中位入住 ICU 时间为 166.917 h，最长入住 ICU 时间为 358.320 h。入选病例的其他基线特征如表 1 所示。将 1 783 例样本按照 9 ∶ 1 的比例随机划分为训练集和测试集，最终训练集样本共 1 604 例，其中死亡病例 235 例（13.18%）；测试集样本共 179 例，其中死亡病例 27 例（1.51%）。

表 1　入组创伤病例的基线资料

Tab 1　Baseline information of trauma patients enrolled in this study

Index	Total N=1 783	Death group N=262	Non-death group N=1 521	Statistic	P value
Gender, n (%)				χ^2=0.399	0.286
Male	904 (50.70)	138 (52.67)	769 (50.56)		
Female	879 (49.30)	124 (47.33)	752 (49.44)		
Race, n (%)				χ^2=5.310	0.150
White	976 (54.74)	132 (13.52)	844 (86.47)		
Black (Brown)	437 (24.51)	67 (15.33)	370 (84.67)		
Asian	151 (8.47)	31 (20.53)	120 (79.47)		
Others	219 (12.28)	32 (14.61)	187 (85.39)		
Age/year, $\bar{x}\pm s$	57.072±7.275	60.785±8.906	56.432±6.753	t=9.154	<0.001
Body height/m, $\bar{x}\pm s$	1.772±0.113	1.763±0.118	1.772±0.111	t=1.057	0.291
Body weight/kg, $\bar{x}\pm s$	79.667±8.590	80.004±8.053	79.609±8.678	t=0.688	0.492
Heart rate/min^{-1}, $\bar{x}\pm s$	89.223±6.724	92.695±7.576	88.625±6.378	t=9.265	<0.001
Respiratory rate/min^{-1}, $\bar{x}\pm s$	15.630±2.860	16.978±3.165	15.398±2.738	t=8.422	<0.001
Body temperature/℃, $\bar{x}\pm s$	36.766±0.662	36.461±0.503	36.818±0.672	t=8.211	<0.001
SBP/mmHg, $\bar{x}\pm s$	109.597±15.154	96.590±10.526	111.837±14.695	t=16.096	<0.001
DBP/mmHg, $\bar{x}\pm s$	76.834±9.795	64.812±6.773	78.905±8.682	t=24.995	<0.001
Hemoglobin/(g•L^{-1}), $\bar{x}\pm s$	97.992±12.990	93.061±14.525	98.846±12.517	t=6.738	<0.001
WBC/(L^{-1}, ×10^9), $\bar{x}\pm s$	10.185±2.482	11.012±2.395	10.072±2.489	t=5.125	<0.001
Platelet/(L^{-1}, ×10^9), $\bar{x}\pm s$	219.939±30.763	216.956±31.078	220.453±30.679	t=1.701	0.089
ALT/(U•L^{-1}), $\bar{x}\pm s$	33.328±6.616	37.768±5.821	32.563±6.442	t=12.245	<0.001
AST/(U•L^{-1}), $\bar{x}\pm s$	30.728±7.464	36.217±7.554	29.783±7.028	t=13.533	<0.001
TBil(μmol•L^{-1}), $\bar{x}\pm s$	11.460±3.150	13.796±2.756	11.210±3.060	t=12.401	<0.001
Scr/(μmol•L^{-1}), $\bar{x}\pm s$	85.914±14.943	82.745±16.797	86.495±14.746	t=4.358	<0.001
BUN/(mmol•L^{-1}), $\bar{x}\pm s$	4.736±0.822	5.018±0.965	4.711±0.799	t=3.359	<0.001
PO_2/mmHg, $\bar{x}\pm s$	96.111±3.583	93.733±4.065	96.521±3.326	t=12.102	<0.001
PCO_2/mmHg, $\bar{x}\pm s$	37.138±3.167	37.669±2.798	37.047±3.217	t=2.944	0.003
SO_2/%, $\bar{x}\pm s$	95.540±2.668	90.176±2.694	96.223±1.745	t=31.361	<0.001
ICU time/h, $\bar{x}\pm s$	162.096±55.268	84.760±80.211	175.416±35.608	t=17.991	<0.001

1 mmHg=0.133 kPa. SBP: Systolic blood pressure; DBP: Diastolic blood pressure; WBC: White blood cell; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; TBil: Total bilirubin; Scr: Serum creatinine; BUN: Blood urea nitrogen; PO_2: Partial pressure of oxygen; PCO_2: Partial pressure of carbon dioxide; SO_2: Saturation of oxygen; ICU: Intensive care unit.

2.2　模型分析变量选择　使用 XGBoost、AdaBoost 和 RF 算法对进入 ICU 后的首次在院数据进行建模并输出各自模型的特征重要性指数，每种算法计算 5 次特征重要性指数后汇总平均数，其中排在前 6 位的变量分别是血红蛋白、收缩压、舒张压、年龄、心率和氧分压（表 2）。这 3 种机器学习算法得到的特征重要性指数的 ICC 为 0.909。在综合了机器分析、相关研究文献及专家评估等多种建议后，将年龄、心率、收缩压、血红蛋白、总胆红素、血肌酐和氧分压等 7 个变量作为训练动态模型的分析变量。

表 2　3 种机器学习模型输出的各变量的特征重要性指数

Tab 2　Feature importance for each variable output from 3 machine learning models

$n=5$

Feature	RF model	XGBoost model	AdaBoost model	Average	Weight/%
Hemoglobin, $\bar{x}\pm s$	0.554±0.039	0.610±0.035	0.613±0.050	0.593	8.474
SBP, $\bar{x}\pm s$	0.534±0.012	0.634±0.012	0.560±0.017	0.576	8.233
DBP, $\bar{x}\pm s$	0.450±0.010	0.483±0.010	0.635±0.019	0.523	7.476
Age, $\bar{x}\pm s$	0.521±0.021	0.552±0.050	0.432±0.016	0.502	7.173
Heart rate, $\bar{x}\pm s$	0.384±0.035	0.302±0.078	0.550±0.009	0.412	5.892
PO_2, $\bar{x}\pm s$	0.434±0.011	0.335±0.014	0.447±0.014	0.406	5.800
Respiratory rate, $\bar{x}\pm s$	0.348±0.013	0.324±0.016	0.437±0.012	0.370	5.291
BUN, $\bar{x}\pm s$	0.350±0.013	0.313±0.007	0.344±0.059	0.336	4.801
WBC, $\bar{x}\pm s$	0.262±0.015	0.314±0.018	0.427±0.010	0.334	4.783
Scr, $\bar{x}\pm s$	0.228±0.018	0.275±0.012	0.424±0.008	0.309	4.422
Body weight, $\bar{x}\pm s$	0.169±0.012	0.416±0.024	0.285±0.018	0.290	4.147
TBil, $\bar{x}\pm s$	0.155±0.013	0.361±0.013	0.354±0.017	0.290	4.143
SO_2, $\bar{x}\pm s$	0.182±0.006	0.349±0.001	0.332±0.022	0.288	4.112
Platelet, $\bar{x}\pm s$	0.161±0.010	0.332±0.010	0.350±0.028	0.281	4.019
PCO_2, $\bar{x}\pm s$	0.172±0.011	0.317±0.017	0.344±0.018	0.278	3.973
Body temperature, $\bar{x}\pm s$	0.139±0.014	0.316±0.020	0.368±0.035	0.275	3.925
Body height, $\bar{x}\pm s$	0.068±0.015	0.327±0.015	0.315±0.011	0.237	3.384
ALT, $\bar{x}\pm s$	0.159±0.029	0.303±0.045	0.245±0.036	0.235	3.367
AST, $\bar{x}\pm s$	0.061±0.010	0.309±0.020	0.284±0.007	0.218	3.118
Gender, $\bar{x}\pm s$	0.048±0.014	0.181±0.018	0.137±0.011	0.122	1.745
Race, $\bar{x}\pm s$	0.029±0.008	0.172±0.014	0.159±0.035	0.120	1.721
ICC (95% *CI*)	0.909 (0.811, 0.960)				

SBP: Systolic blood pressure; DBP: Diastolic blood pressure; PO_2: Partial pressure of oxygen; BUN: Blood urea nitrogen; WBC: White blood cell; Scr: Serum creatinine; TBil: Total bilirubin; SO_2: Saturation of oxygen; PCO_2: Partial pressure of carbon dioxide; ALT: Alanine aminotransferase; AST: Aspartate aminotransferase; ICC: Intraclass correlation coefficient; *CI*: Confidence interval.

2.3　模型预测性能　在时间窗分别为 4、6、8 h 时使用 LSTM 和 GRU 算法训练的动态模型预测临床结局（死亡与否）的灵敏度、特异度、F1 值和 AUC 值见表 3。在 8 h 时间窗下，LSTM 和 GRU 模型各节点的平均特异度分别为 0.912±0.025 和 0.910±0.034，平均灵敏度分别为 0.814±0.044 和 0.813±0.026。而时间窗口为 6 h 和 4 h 时 2 种算法模型的预测性能均不及时间窗口为 8 h 时。

表 3　2 种 RNN 动态模型的各节点预测性能

Tab 3　Prediction performance of 2 RNN dynamic models under different time windows

$n=5, \bar{x}\pm s$

Time window	Model	Sensitivity	Specificity	F1 value	AUC
4 h	LSTM	0.757±0.042	0.847±0.029	0.510±0.053	0.767±0.034
	GRU	0.742±0.028	0.841±0.020	0.513±0.048	0.752±0.035
6 h	LSTM	0.782±0.043	0.879±0.025	0.525±0.055	0.792±0.032
	GRU	0.780±0.026	0.872±0.014	0.530±0.050	0.792±0.036
8 h	LSTM	0.814±0.044	0.912±0.025	0.547±0.056	0.826±0.034
	GRU	0.813±0.026	0.910±0.034	0.552±0.053	0.825±0.037

RNN: Recurrent neural network; AUC: Area under curve; LSTM: Long short-term memory; GRU: Gated recurrent unit.

RNN 算法模型的 4 个性能指标的全因子差异性分析结果见表 4。在不同时间窗时，RNN 算法模型预测临床结局的灵敏度、特异度、F1 值和 AUC 值差异均有统计学意义（均 $P<0.001$）；在不同 RNN 算法（LSTM 和 GRU）间仅特异度差异有统计学意义（$P=0.036$），而灵敏度、F1 值和 AUC 值在不同 RNN 算法间及全部 4 个性能指标的算法与时间窗交互作用均无统计学意义（均 $P>0.05$）。

基于 2 种 RNN 算法训练的预测模型在 8 h 时间窗时的性能表现见表 5。基于 LSTM、GRU、HMM、RF 和 logistic 算法的预测模型在各节点的平均 AUC 值分别为 0.826±0.034、0.825±0.037、0.742±0.015、0.707±0.019 和 0.644±0.033。各模型的灵敏度、特异度、F1 值和 AUC 值差异均有统计学意义（均 $P<0.001$），其中基于 LSTM、GRU 和 HMM 算法的动态预测模型的各性能指标均高于 RF 和 logistic 模型，且基于 logistic 算法的预测模型各性能指标在 5 种预测模型均最低。在一致性方面，5 种预测模型在灵敏度、特异度、F1 值方面的 ICC 分别为 0.262、0.244、0.395，而 AUC 值的 ICC 仅 0.002。各算法模型性能指标的时间趋势变化见图 2。

表 4　不同影响因素下 RNN 模型性能分析

Tab 4　Analysis of performance of RNN models under different influencing factors

Factor	Evaluation index	EMS	*F* value	*P* value
Time window	Sensitivity	0.169	130.931	＜0.001
	Specificity	0.184	285.298	＜0.001
	F1 value	0.061	21.681	＜0.001
	AUC	0.181	149.970	＜0.001
RNN algorithm	Sensitivity	0.005	3.603	0.058
	Specificity	0.003	4.409	0.036
	F1 value	0.002	0.800	0.372
	AUC	0.003	2.681	0.102
Time window-RNN algorithm	Sensitivity	0.002	1.768	0.172
	Specificity	＜0.001	0.283	0.754
	F1 value	＜0.001	0.008	0.992
	AUC	0.003	2.693	0.069

RNN: Recurrent neural network; EMS: Effect mean square; AUC: Area under curve.

表 5　不同算法模型预测性能的差异性及一致性分析

Tab 5　Analysis of difference and consistency of prediction performance of different algorithm models

Evaluation index	Model	Average value, $\bar{x}\pm s$	*F* value	*P* value	ICC (95% *CI*)
Sensitivity	LSTM	0.814±0.044	750.400	＜0.001	0.262 (0.167, 0.374)
	GRU	0.813±0.026			
	HMM	0.780±0.011			
	RF	0.725±0.062			
	Logistic	0.645±0.082			
Specificity	LSTM	0.912±0.025	777.767	＜0.001	0.244 (0.143, 0.355)
	GRU	0.910±0.034			
	HMM	0.871±0.013			
	RF	0.801±0.026			
	Logistic	0.705±0.038			
F1 value	LSTM	0.547±0.056	332.727	＜0.001	0.395 (0.313, 0.525)
	GRU	0.552±0.053			
	HMM	0.496±0.053			
	RF	0.463±0.032			
	Logistic	0.327±0.033			
AUC	LSTM	0.826±0.034	606.155	＜0.001	0.002 (−0.058, 0.082)
	GRU	0.825±0.037			
	HMM	0.742±0.015			
	RF	0.707±0.019			
	Logistic	0.644±0.033			

ICC: Intraclass correlation coefficient; *CI*: Confidence interval; LSTM: Long short-term memory; GRU: Gated recurrent unit; HMM: Hidden Markov model; RF: Random forest; AUC: Area under curve.

3　讨　论

本研究提出了基于 RNN 算法对创伤重症患者转归进行实时预测的动态模型，2 种 RNN 算法模型的平均 AUC 值最高可到 0.826±0.034（8 h 时间窗），而最低也在 0.752±0.035（4 h 时间窗），而灵敏度和特异度最低为 0.742±0.028 和 0.841±0.020，说明本研究提出的动态模型具有较好的预测效果。

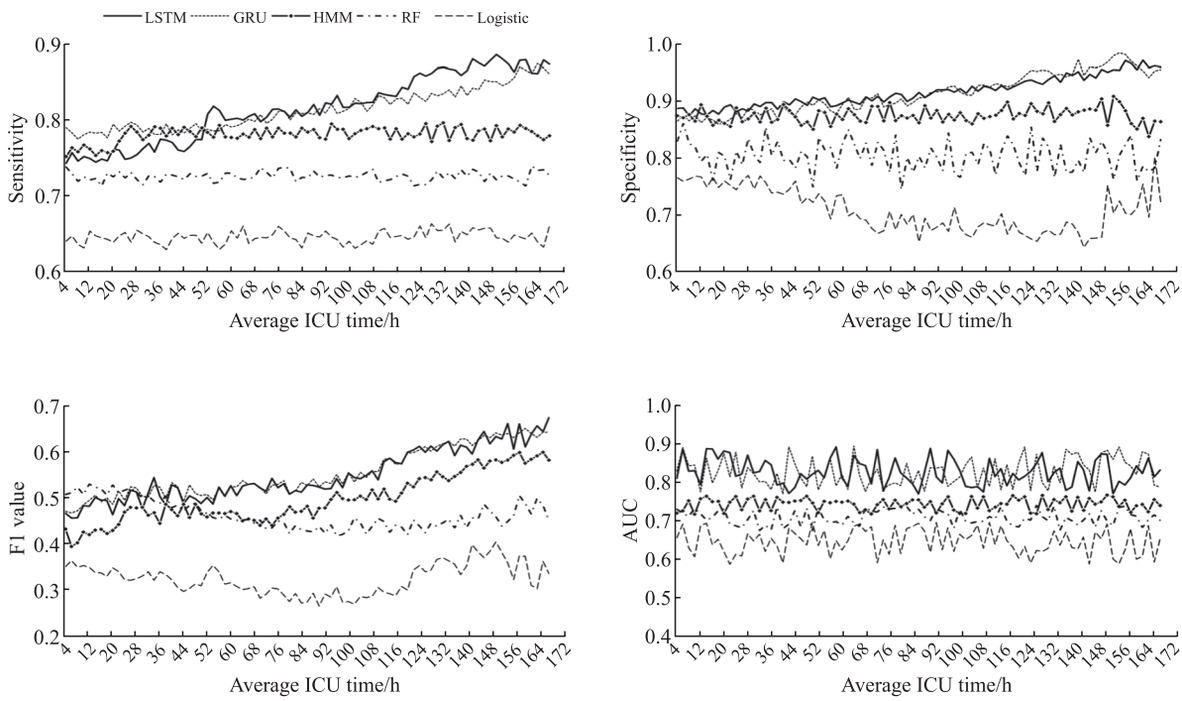

图 2　各算法模型预测性能的时间趋势图

Fig 2　Time trend plot of predictive performance of each model

LSTM: Long short-term memory; GRU: Gated recurrent unit; HMM: Hidden Markov model; RF: Random forest; ICU: Intensive care unit; AUC: Area under curve.

本研究结果显示，各模型在 8 h 时间窗口下的预测性能指标均相对优于 6 h 和 4 h，6 h 和 4 h 时各项指标性能分别约为 8 h 时的 95% 和 92%，说明时间窗长度可能会对 RNN 动态模型的预测性能造成影响，可能原因是时间窗的增长令模型在每个循环节点上所使用的时间序列数据更多，捕捉时间连续变化信息也更准确[20]。此外，尽管本研究中不同 RNN 算法（LSTM 和 GRU）模型的性能指标中仅特异度差异有统计学意义（P＝0.036），但有报道指出 LSTM 算法相较于 GRU 算法内部更精密（参数更多），在大数据长时程的预测场景中 LSTM 算法的预测准确性可能高于 GRU 算法，但 LSTM 算法所需要调用的计算资源和时间成本明显高于 GRU 算法[21]。有研究报道，在同等条件下 GRU 算法的训练收敛时间和单次遍历时间都明显小于 LSTM 算法[22]，相比之下 GRU 算法对于硬件条件有限的环境更加友好。

在对不同算法模型的横向比较中，基于 LSTM、GRU 和 HMM 算法的动态预测模型的各项性能指标均高于 RF 和 logistic 模型，这表明动态预测策略较静态预测策略或许更具优势。LSTM、GRU 和 HMM 算法可以使用时间序列数据进行预测，RF 和 logistic 模型仅能分析横截面数据，而时间序列数据显然比横截面数据能更加全面、详实地反映疾病情况。同为动态模型的 RNN 和 HMM 算法，前者的预测性能优于后者。就 AUC 值而言，LSTM 和 GRU 算法较 HMM 算法总体提升了约 11%；其他性能指标同样显著提升。这一结果表明，在相同的动态策略加持下，RNN 算法本身的复杂而精密优势得以体现。有研究表明，HMM 算法内部是线性连接，故其对于非线性变化的拟合明显弱于 RNN 算法[23]。

本研究中，灵敏度、特异度和 F1 值在各模型间的一致性均较小（ICC＜0.400），而 AUC 值一致性在统计学上证据不足（95% *CI* 包含 0），说明各算法模型性能随时间推移其变化波动分歧较为明显。从各算法模型性能指标的时间趋势图可见，LSTM、GRU 和 HMM 算法模型的灵敏度、特异度和 F1 值时间推移呈上升趋势，而 RF 和 logistic 模型则未见这一现象，这表明动态预测模型或许可以更有效地利用分析变量在时间层面上的连续性，对于病情发展的拟合较静态模型亦或更为精准。值得一提的是，随着循环的进行不断有样本从队列脱落（死亡或好转），尤其是在后期节点样本量明显

下降，而 RNN 动态模型的预测性能依旧可以保持相对稳定，本研究认为循环节点间模型训练的参数“继承”策略或许在其中发挥了一定作用[12]。

本研究对于变量筛选的原则是特征重要性尽量高，但相关性尽量小，其目的是保留特征贡献的同时抑制模型的内生性和共线性。从这个角度出发，本研究选择分析变量时并未完全依据特征重要性指数排序，而是采纳了部分文献报道和专家建议。本研究在征求临床专家有关分析变量建议时，反馈的结果认为失血性休克及多器官功能衰竭是造成创伤患者死亡的关键因素之一，而部分研究报道[24-25]也支持了该观点，故采纳的分析变量多与人体主要脏器功能关联紧密。从预测结果的角度看，本研究 RNN 算法模型采用年龄、心率、收缩压、血红蛋白、总胆红素、血肌酐和氧分压等 7 个分析变量，达到了良好的预测效果，说明本研究采用的分析变量组合是合理且可被接受的。

从 PubMed 和中国知网检索到在近 10 年（2013－2022 年）公开报道的各类创伤重症相关预测研究有 3 万余篇，但是，由于不同研究之间在包括策略架构、模型算法、样本量、预测结局、评价指标等在内的诸多因素差异较大，研究和研究之间、模型和模型之间可比性并不高。必须承认的是，本研究所训练的 RNN 模型 F1 值并不理想，各节点平均 F1 值最高也仅为 0.552±0.053（8 h 时间窗）。但笔者认为本研究中 RNN 模型 F1 值较低主要是由正负例样本不均衡造成的。从所采集的数据可以发现，本研究中死亡与非死亡病例比例仅为 1 ∶ 5.805，这会使被错误分类的负例样本（非死亡病例）极大地干扰 F1 值，因此可以认为 F1 值较低并不能说明 RNN 模型的性能不理想，还需要结合更大样本的数据进行全面评价。有报道指出，在错分代价较为敏感的应用场景，AUC 值对于模型性能评价可能更具优势[26]；另有研究认为当模型的 AUC 值达到 0.8 以上时，其已可接受为较为理想的分类器模型[27]。

本研究存在一定局限性。首先，本研究使用的样本数据略显不足，且为单中心数据；其次，RNN 模型的可解释性欠缺，后期可考虑通过引入局部可解释的模型无关解释算法（local interpretable model-agnostic explanations）[28]或注意力机制[29]等方法尝试弥补该缺失。

综上所述，本研究在创伤重症背景下成功搭建了基于 RNN 算法的动态预测模型并将其应用于创伤重症患者的临床结局预测，初步论证了该动态预测模型搭建方案的可行性和实现路径，可为继续开展 RNN 模型临床应用研究提供参考。

[参考文献]

[1] 刘国辉. 重症创伤患者的一体化救治模式[J]. 中华急诊医学杂志，2013，22(6)：569-570. DOI: 10.3760/cma.j.issn.1671-0282.2013.06.003.

[2] 田振中，孙振雷. 我国道路交通事故死亡人数影响因素及管理对策研究[J]. 中国人民公安大学学报(自然科学版)，2022，28(2)：38-44. DOI: 10.3969/j.issn.1007-1784.2022.02.006.

[3] 张思森，岳茂兴，王立祥. 创伤性休克急救复苏新技术临床应用中国专家共识(2019)[J]. 中华卫生应急电子杂志，2019，5(1)：1-6. DOI：10.3877 /cma.j.issn.2095-9133.2019.01.001.

[4] 陈逍堃，林维成，张鹏，等. 创伤急救体系的发展与现状[J]. 北京大学学报(医学版)，2017，49(2)：368-371. DOI: 10.3969/j.issn.1671-167X.2017.02.034.

[5] BARAK O. Recurrent neural networks as versatile tools of neuroscience research[J]. Curr Opin Neurobiol, 2017, 46: 1-6. DOI: 10.1016/j.conb.2017.06.003.

[6] KRIEGESKORTE N, GOLAN T. Neural network models and deep learning[J]. Curr Biol, 2019, 29(7): R231-R236. DOI: 10.1016/j.cub.2019.02.034.

[7] 张家艳，郑建立，郑西川，等. MIMIC 数据库智能挖掘研究概述[J]. 计算机技术与发展，2020，30(1)：144-148. DOI: 10.3969/j.issn.1673-629X.2020.01.026.

[8] 胡中源，薛羽，查加杰. 演化循环神经网络研究综述[J]. 计算机科学，2023，50(3)：254-265. DOI: 10.11896/jsjkx.220600007.

[9] KWON B C, CHOI M J, KIM J T, et al. RetainVis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records[J]. IEEE Trans Vis Comput Graph, 2018: 299-309. DOI: 10.1109/TVCG.2018.2865027.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Comput, 1997, 9(8): 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

[11] GRAVES A, FERNÁNDEZ S, SCHMIDHUBER J. Bidirectional LSTM networks for improved phoneme classification and recognition[M]//Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 799-804. DOI: 10.1007/11550907_126.

[12] YU Z, SHEN D, JIN Z, et al. Progressive transfer learning[J]. IEEE Trans Image Process, 2022, 31: 1340-1348. DOI: 10.1109/TIP.2022.3141258.

[13] 王成，刘亚峰，王新成，等. 分类器的分类性能评价指

标[J]. 电子设计工程，2011，19(8)：13-15，21. DOI: 10.3969/j.issn.1674-6236.2011.08.004.

[14] 刘伟平，黄晨浩. 基于AUC的支持向量机分类方法及应用研究[J]. 湖南城市学院学报(自然科学版)，2023，32(6)：69-73. DOI: 10.3969/j.issn.1672-7304.2023.06.0012.

[15] RESÉNDIZ ROJAS M, FONTECAVE-JALLON J, RIVET B. Hidden Markov model in nonnegative matrix factorization for fetal heart rate estimation using physiological priors[J]. Physiol Meas, 2022, 43(10). DOI: 10.1088/1361-6579/ac92bf.

[16] 董红瑶，王弈丹，李丽红. 随机森林优化算法综述[J]. 信息与电脑，2021，33(17)：34-37. DOI: 10.3969/j.issn.1003-9767.2021.17.011.

[17] 齐巧娜，刘艳，陈霁晖，等. 机器学习XGBoost算法在医学领域的应用研究进展[J]. 分子影像学杂志，2021，44(5)：856-862. DOI: 10.12122/j.issn.1674-4500.2021.05.25.

[18] 徐洪学，孙万有，杜英魁，等. 机器学习经典算法及其应用研究综述[J]. 电脑知识与技术，2020，16(33)：17-19.

[19] 余红梅，罗艳虹，萨建，等. 组内相关系数及其软件实现[J]. 中国卫生统计，2011，28(5)：497-500. DOI: 10.3969/j.issn.1002-3674.2011.05.006.

[20] MAJUMDAR A, GUPTA M. Recurrent transform learning[J]. Neural Netw, 2019, 118: 271-279. DOI: 10.1016/j.neunet.2019.07.003.

[21] 邱锡鹏. 神经网络与深度学习[M]. 北京：机械工业出版社，2020：141-144.

[22] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. 2014: arXiv: 1412.3555 (2014-12-11)[2024-03-21]. *http://arxiv.org/abs/1412.3555.*

[23] YU S Z. Explicit duration recurrent networks[J]. IEEE Trans Neural Netw Learn Syst, 2022, 33(7): 3120-3130. DOI: 10.1109/TNNLS.2021.3051019.

[24] 陈静清，周练兴，卢善儒，等. 68 例急诊创伤死亡病例分析及临床意义研究[J]. 现代诊断与治疗，2014，25(20)：4778-4779.

[25] 王占科，胡新勇，柴长春，等. 357 例创伤死亡患者空腹血糖与多器官功能不全综合征相关分析[J]. 现代诊断与治疗，2005，16(2)：72-74. DOI: 10.3969/j.issn.1001-8174.2005.02.004.

[26] 汪云云，陈松灿. 基于AUC的分类器评价和设计综述[J]. 模式识别与人工智能，2011，24(1)：64-71. DOI: 10.3969/j.issn.1003-6059.2011.01.008.

[27] HAND D J, TILL R J. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. Mach Learn, 2001, 45(2): 171-186. DOI: 10.1023/A: 1010920819831.

[28] 林志萍，杨立洪. 基于LIME的改进机器学习可解释性方法[J]. 数据挖掘，2021，11(2)：38-49. DOI: 10.12677/HJDM.2021.112005.

[29] 朱张莉，饶元，吴渊，等. 注意力机制在深度学习中的研究进展[J]. 中文信息学报，2019，33(6)：1-11. DOI: 10.3969/j.issn.1003-0077.2019.06.001.

［本文编辑］　杨亚红