

Predicting Critically Ill Patients Short-Term Mortality Risk Using Routinely Collected Data: Deep Learning Model Development, Validation, and Explanation

Shangping ZHAO

Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410003, China; Department of Critical Care Medicine, the Third Xiangya Hospital, Central South University, Changsha 410013, China
E-mail: chinazsp@163.com

Pan LIU

Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410003, China
E-mail: liupan09@nudt.edu.cn

Guohui LI

Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410003, China
E-mail: guohli@nudt.edu.cn

Yanming GUO

Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410003, China
E-mail: guoyanming@nudt.edu.cn

Guanxiu TANG*

Nursing Department, the Third Xiangya Hospital, Central South University, Changsha 410013, China
E-mail: tangguanxiu@163.com

Abstract This paper aims to develop and validate a deep learning-based short-term mortality risk prediction model for critically ill patients by using routinely collected data in a large Chinese cohort and explore the explainability of the model decision. A total of 10925 critically ill patients between January 2014 and June 2020 are included in this study. Data routinely collected in the electronic health records (EHRs) system are extracted and used to develop a short-term mortality risk prediction model based on a deep artificial neural network (ANN). The features include demographic characteristics, vital signs, laboratory tests, and the daily dose of intravenous medications. The developed deep learning model (AUROC: 0.88, AUPRC: 0.63, Brier score: 0.108) is superior to the model based on APACHE II scores (AUROC: 0.78, AURPC: 0.52, Brier score: 0.124) in the prediction of hospital mortality for

Received September 26, 2022, accepted March 19, 2023
Supported by Xiangya Clinical Big Data Construction Project
*Corresponding author

critically ill patients. Further attribution analysis based on the integrated gradients method shows that measurements observed at a later time seem to have a more significant influence on mortality, while earlier usage of amiodarone or dexmedetomidine contributed to lower mortality. This well-performing and interpretable model may have practical implications for improving the quality of care for critically ill patients.

Keywords Critically ill patients; deep learning; artificial neural network; mortality; prediction; APACHE II

1 Introduction

Reliable and real-time mortality predictions for critically ill patients are essential for assessing the severity of illness and determining the value of novel interventions, which may help to promote the quality of care and to improve clinical outcomes^[1]. It is very challenging for health care professionals to respond in a timely manner to the massive amounts of data generated faster than the human brain can interpret. Therefore, considerable efforts have been invested in predicting the risk of death in critically ill patients, and several severity scoring systems have been developed. The most documented mortality risk prediction tools include the Acute Physiology and Chronic Health Evaluation (APACHE), the Mortality Probability Model (MPM), and the Simplified Acute Physiology Score (SAPS)^[2, 3]. Estimates suggest, however, that only 12% of intensive care units (ICUs) use these mortality risk assessment tools^[4]. Low levels of adoption are mainly attributed to the time- and labor-consuming collection of patient data that is not captured in routine clinical workflows^[5] and the relatively poor accuracy of traditional predictive tools with limited predictors involved in the models^[6].

Recently, advanced machine learning (ML) techniques, such as deep learning, have shown great promise in various prediction tasks and have allowed advanced, accurate predictions of mortality based on high-dimensional, confounding medical data extracted from electronic health records (EHRs). In a study by Alexander, a predictive tool based on recurrent neural networks was developed to help identify high-risk patients with severe complications in postcardiosurgical care^[7], and the ML method achieved an AUROC that significantly surpassed that of clinical reference tools. Similarly, in a study^[8] from 2019, Holmgren found that an ANN-based model outperformed the SAPS 3 model in the early prediction of 30-day mortality in ICU patients while using the same predictors. More interestingly, Alghatani applied ML binary classification algorithms to predict hospital mortality for ICU patients and achieved a reasonable performance based on demographic characteristics and vital signs^[9]. These results are encouraging and can help direct the attention of medical staff members toward patients who are most at risk.

Although increasing numbers of studies are using ML to analyze ICU data, ML-based mortality prediction models for patients have not been widely applied in clinical practice^[10]. Most of these mortality models are developed based on public datasets, such as the Medical Information Mart for Intensive Care (MIMIC) database and the eICU collaborative research database^[11, 12], and very few have contributed meaningfully to clinical care^[13]. It should also be noted that some of the studies validated their predictions using random subsets of the development data, which may have resulted in the overestimation of model performance^[14]. Without external validation, the performance of models evaluated in those studies in Chinese hospitals is un-

clear, considering the differences in demographic characteristics, medical resource allocation, and technical level among different regions. Additionally, many researchers focus mainly on optimizing the performance of the models while disregarding their explainability, which could help professionals better understand the reasons for the model decisions^[15]. Attempts to develop, validate and interpret prediction models based on ML algorithms by using routinely collected clinical EHR data in real-world settings may potentially increase the use of these models in clinical practice.

In this study, we aimed to develop and validate a deep learning-based short-term mortality risk prediction model by using routinely collected data in clinical practice in a large Chinese real-world hospital setting. Our work contributes to the literature through the development of a well-performing and interpretable mortality risk prediction model based on deep neural networks that uses easily accessible variables derived from a 3-day window prior to patient discharge as features. Additionally, attribution analysis highlights the strong predictive role of variables observed at a later time in mortality risk prediction, revealing that the early use of amiodarone and dexmedetomidine when needed may be related to a lower mortality risk. The proposed model may facilitate the improvement of quality of care for critically ill patients.

2 Methods

2.1 Patient Selection and Data Sources

Critically ill patients over the age of 14 who were discharged from a general hospital in Hunan, China between January 2014 and June 2020 were included in this study. All data were obtained from the Intensive Care Database of the hospital. Critically ill patients were defined as those who once had a record of “critically ill status” during hospitalization, as found in the physician’s order. Therefore, we included both ICU and non-ICU inpatients. As vital signs and physiological laboratory variables measured at different admissions for the same patient could be highly correlated, repeated sampling may overestimate the predictive performance of the mortality model; we thus only included the last hospitalization record if the patients had several hospitalizations. Subjects who stayed less than 24 hours in the hospital were also excluded to ensure sufficient data for analysis. This study was approved by the institutional ethics committee of the hospital (2020-S626).

Except for those who had a record of death in the hospital information system (HIS), patients who were in an extremely unstable physiological state at discharge despite having received active therapeutic interventions to support life were also defined as positive samples if their families gave up treatment and ask that the patient be sent back home. To label these positive samples, we invited two professional clinicians to check and review the patients’ discharge records (which record the patients’ current physiological state, treatments and reason for discharge) to identify patients who had no death records in the HIS but were actually discharged with a poor prognosis. According to our clinical experience, these patients were highly likely to die soon after discharge from the hospital (99% of patients died within 24 hours postdischarge in our previous follow-up) and therefore were defined as positive samples in this study. This definition of positive samples may reduce the underestimation of mortality in the real setting, as the above phenomenon is common in China.

2.2 Feature Selection

Clinical data were captured as part of the usual care processes in the EHR system and archived in the Intensive Care Database of the hospital. We attempted to extract all of the available variables, and the overview of the features is shown in Table 1. Data were derived from four primary sources: 1) demographic data, including age and sex; 2) vital signs; 3) physiological laboratory test results (e.g., blood counts, arterial blood gas, and electrolytes); and 4) daily dose of the intravenous drug. For the physiological laboratory test variables, all of the subitems included in the tests were used as predictors. Detailed information on the features is described in our previous study^[16].

Table 1 Features in the deep learning model overview

Features category	Variables
Demographic data	Age, gender
Vital signs	Temperature, Heart rate, Respiratory rate, Systolic blood pressure, Diastolic blood pressure, Oxygen saturation
Laboratory test results (including eight laboratory tests)	Blood routine results, Clotting function, Liver function, Kidney function, Electrolyte, Myocardial enzyme, Glucose and blood lipid, Arterial blood gas
Intravenous medications per day (including 54 drugs)	Vasoactive drugs (7 drugs), Positive inotropic drugs (6 drugs), Antibiotic drugs (15 drugs), Anaesthetic drugs (10 drugs), Nutrition-related drugs (5 drugs), Hormone (5 drugs), Blood products (6 drugs)

To better discriminate deceased patients from the whole sample, we focused more on the time when the patients were relieved from critically ill status, which was indicated by a health status mark provided by the doctors in the physician order. We defined this time point as t_0 . The patient either became relieved from the status or died at this time point, when more significant differences may exist in clinical characteristics between positive and negative samples. D_1 was defined as the period of the day on t_0 but no later than t_0 ; D_2 was the day before D_1 , and D_3 was the day before D_2 . Considering that the variation in clinical data may be more sensitive to patients' clinical outcomes, we used multidimensional data collected during a relatively long period for mortality prediction instead of that collected within 24 hours or less. The median length of ICU stay of the subjects was 2.5 days, and therefore, we extracted time-series data in reverse chronological order from D_1 to D_3 . For vital signs and laboratory test variables, the maximum, minimum, and average values for each day were calculated and used as predictors. Attributes with coverage below 10% were not used for analysis, leaving 813 attributes for mortality prediction.

2.3 Data Preprocessing

Extreme and error values failing the logic check were censored. To address outliers, we analyzed the distribution of the numerical variables, and it was reviewed by critical physicians. For physiological variables, such as vital signs and laboratory test results, we capped the extreme values at the 1st and 99th percentiles. After that, an additional list of refined clinically reason-

able variable ranges was established based on the clinical experts' knowledge of valid clinical measure ranges. Each variable was also associated with lower and upper thresholds for defining a physiologically valid range. Any value that fell outside the physiologically valid range was treated as a missing value.

Missing values of physiological measurements at a certain timestep was imputed by the observed value at the nearest forward time point. If the variable was never observed for a patient, it was set to a prespecified normal value. We used this imputed scheme based on two considerations. First, doctors always refer to the latest medical measurements to help judge the severity of illness. Second, if the variable was never observed for a patient during hospitalization, it is highly likely that the variable had no effect on the prognosis of the patient; thus, there was no need to measure it. The value of the daily drug dose was recorded as zero if there was no prescription drug record during the corresponding period.

2.4 Model Development and Evaluation

To compare the performance of the deep learning model in predicting mortality with the model based on the APACHE II scores, 689 cases with APACHE II score records in the EHRs were retained as the test set. The remaining cases were used for model development, with 90% of the data as the training set and 10% as the validation set. The APACHE II scores were used to calculate the risk of death for patients according to the formula described in a previous study^[3].

We employed the SoftMax logistic loss function at the output layer. Considering the imbalanced nature of the sample, the focal loss function was used, and it was optimized using the Adam implementation of batch gradient descent, with a learning rate of 0.001. To prevent overfitting, we set a 10% dropout rate for hidden layers. Intermediate model snapshots were taken every 100 mini-batch iterations, and the model that performed best in the validation set was selected as the final model.

To select an appropriate network, the network configuration was reached by an extensive grid search^[17] over various network depths, nodes, decay rate and activation functions (Tanh, ReLU and SeLU). We constructed multiple deep neural networks using between 2 and 18 hidden layers, where the number of nodes in each layer was log-sampled between 64 and 512, and performed regularization using log-sampled weight decay with the decay parameter λ , ranging from 0.1 to 10^[18]. The best performed model evaluated by area under precision-recall curve (AUPRC), a measure that combines recall and precision for ranked retrieval results^[19], on the validation test was selected as the final model. Then, we obtained the hyperparameters of our final model: 18 hidden layers with 256 nodes in each layer, a weight decay of $\lambda = 0.1$ and ReLU as the activation function. These networks were constructed using the Python programming language (version 3.7.7), the PyTorch framework, and the scikit-learn library (version 0.17.1). The training was performed on an NVIDIA TitanX (12 GB RAM) with CUDA version 8.0.

Model performance measurement and the reported scores were based on the results in the test set, providing an unbiased estimate of model performance. Model discrimination was assessed via the AUROC^[20]. Since the sample was imbalanced (19.01% mortality prevalence), we assessed accuracy using the AUPRC. Furthermore, we computed the Brier score, which measures the calibration of a set of probabilistic predictions, and a lower Brier score indicates a

superior model^[21]. Patient characteristics are presented as the mean (standard deviation, SD), median (interquartile range, IQR), or number (percentage). We checked the characteristics of patients in different subsets. Data were compared using *t* tests, chi-squared tests or Mann-Whitney U tests based on the type and distribution of the data.

2.5 Interpretation of the Deep Learning Model

It is essential to establish the users trust in the model's decisions for them to feel comfortable to take appropriate actions based on those decisions. Providing explanations along with the decisions may help establish this trust. In this study, we used an attribution method called integrated gradients^[22], which has been widely applied to various deep neural network tasks^[23], to analyze the impact of different features on the prediction results. We aimed to rank the features according to the influence evaluated by the method of integrated gradients. For each intravenous drug, we removed all occurrences from the dataset, created a new feature vector, and measured the drop in probability compared to the original probability. This drop in probability is considered the influence the variable had on the model's decision for that patient. Other features were handled as follows. We replaced the age and laboratory test variables with the mean value, swapped the sex from male to female and vice versa, and then measured the respective drops in probability. Finally, the features were sorted according to the strength of the contribution of the features to mortality.

3 Results

3.1 Patient Characteristics

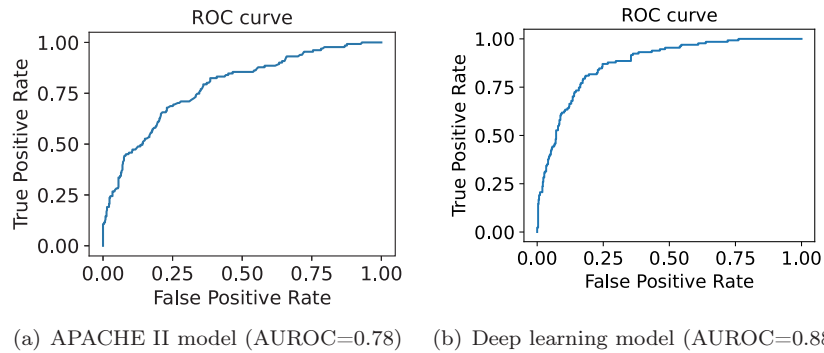
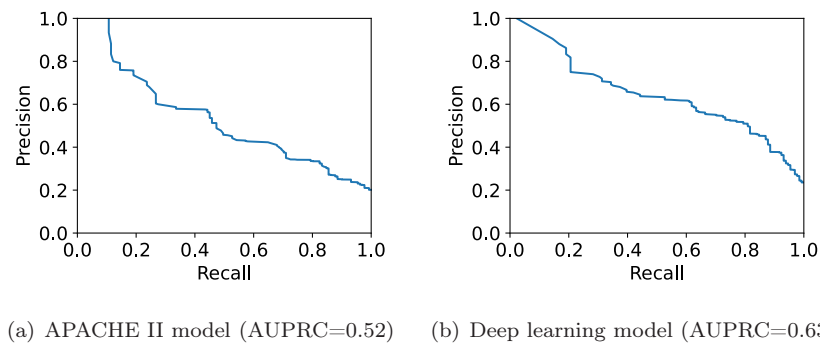
A total of 10,925 critically ill patients were identified in this study, with an average age of 58.89 years and a median hospital length of stay of 12 days (IQR: 7-20), including 3,128 female patients (28.6%) and 9,056 ICU admissions (82.9%). The median length of stay in the ICU was 66 hours (IQR: 13-151), and the overall hospital mortality was 19.0% for the study period. The general characteristics of the patients in the training, validation and testing sets are shown in Table 2. There were no significant differences in age, sex, mortality, ICU admission, emergency admission or comorbidities among the three sets ($P < 0.05$).

3.2 Comparisons of Model Discrimination and Calibration

The performance of the model was calculated based on the patients' data in the independent test set. In terms of discrimination, the AUROC for the deep learning model was 0.88 (95% CI: 0.84-0.91), and that for the APACHE II model was 0.78 (95% CI: 0.73-0.83). The ROC curves for hospital mortality prediction are provided in Figure 1. The precision-recall curves are shown in Figure 2. The deep learning model achieved an AUPRC of 0.63 (95% CI: 0.53-0.71), which was higher than that of the APACHE II model, with an AUPRC of 0.52 (95% CI: 0.44, 0.60). Both the ROC curve and precision-recall curve plots suggested that the deep learning model demonstrated better discrimination performance. To evaluate the agreement between the observed and expected mortality across risk strata, we used the Brier score. The prediction based on the deep learning model exhibited better calibration than that based on the APACHE II model, as reflected by a low Brier score of 0.108 (95% CI: 0.09-0.12). Calibration plots for the two models are shown in Figure 3.

Table 2 Characteristics for patients in the training and testing datasets

	Traning set	Validation set	Testing test	<i>P</i> value
Samples (<i>n</i>)	9212	1024	689	
Age (yr), mean (SD)	58.8 (16.4)	59.0 (16.3)	59.2 (16.1)	0.426
Female sex, <i>n</i> (%)	2625 (28.5)	301 (29.4)	202 (29.3)	0.469
Mortality, <i>n</i> (%)	1741 (18.9)	204 (20.0)	131 (19.0)	0.687
Length of hospital stay, Median days (IQR)	12 (7-20)	12 (7-21)	14 (8-24)	≤ 0.001
ICU admission, <i>n</i> (%)	7195 (78.1)	783 (76.5)	560 (81.3)	0.059
Length of ICU stay, Median hours (IQR)	65 (12-148)	65 (13-140)	88 (24-193)	≤ 0.001
Emergency admition, <i>n</i> (%)	6586 (71.5)	699 (68.3)	498 (72.3)	0.078
Comorbidities, <i>n</i> (%)				
Hypertension	3482(37.8)	391 (38.2)	266 (38.7)	0.866
Diabetes	1621(17.6)	172 (16.8)	117 (17.1)	0.788
Coronary heart disease	1391 (15.1)	151 (14.8)	100 (14.6)	0.935
Stroke	1041(11.3)	10 (0.96)	7 (0.98)	0.150

**Figure 1** Receiver Operating Characteristic (ROC) curves for predicting mortality in critically ill patients**Figure 2** Precision-recall curves for predicting mortality in critically ill patients

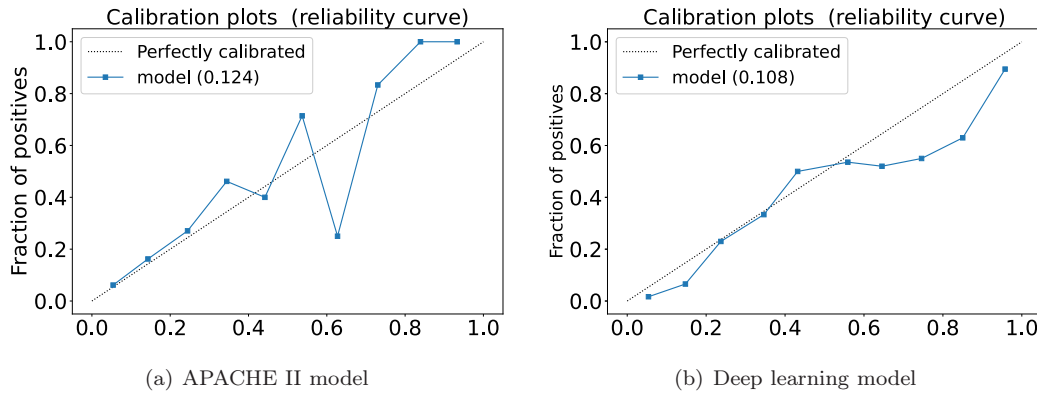


Figure 3 Reliability curves (calibration plots) of the models output probabilities on the test set data

3.3 Explaining the Predictions

The results of attribution analysis using integrated gradients for the deep learning model showed that features measured at D_1 had a greater influence on mortality than features measured at other time points. Among the top 10 ranking features, 6 features were obtained at time point D_1 , and all 10 features were related to the daily dose of medications. The doses of amiodarone and dexmedetomidine used on D_3 had a negative effect on mortality, while the other features had a positive effect on mortality. The ranking of the top 10 features is listed in Table 3.

Table 3 The top 10 ranking of features in the deep learning model

Feature name	Total deviation ^a	Unit contribution ^b	Unit Ranking
Dopamine on D_1	5340	0.483830658	1
Amiodarone on D_3	1200	-0.461598277	2
Noradrenaline on D_1	8557	0.297342469	3
Dexmedetomidine on D_3	3600	-0.248628504	4
Normal saline on D_1	521567	0.207462532	5
Enteral nutrition on D_1	24000	0.198485323	6
Esmolol on D_1	4000	0.197279111	7
Noradrenaline on D_2	4677	0.14894327	8
Teicoplanin on D_1	6200	0.134681246	9
Esmolol on D_3	6100	-0.132296073	10

^a Deviation refers to the difference between the actual value and the baseline value of the feature.

^b The unit contribution is the total contribution (e.g., the sum of the attribution score) divided by the total deviation (e.g., the sum of the deviation). The magnitude of the attribution score signifies the strength of the contribution.

4 Discussion

4.1 Principal Results

To our knowledge, our study is one of only a few attempts to predict in-hospital mortality for critically ill patients by using routinely collected clinical data in a Chinese real-world hospital setting. We used an independent test set to validate the model performance rather than a subset of the developmental data, which may have enhanced the reliability of the validation results. The results showed that the performance of the deep learning model based on our specific data preprocessing approach outperformed the APACHE II model both in discrimination and calibration. In particular, the model reached an AUROC of 0.88, which is not commonly observed in current clinical prognostic models.

Specifically, the prediction was based on time series features collected in reverse chronological order before the specific time when the patients were relieved from critically ill status, rather than variables at a single point within a short time (e.g., 24 h) after ICU admission^[3, 24]. The definition of t_0 is of great significance in clinical practice. It is the time threshold that indicates when the patients either are relieved of their status or have died. The differences in clinical data between the positive and negative samples seem evident at this time point. Deep learning is a form of representation learning in which a machine is fed with raw data and develops its own representations needed for pattern recognition^[25]. When using the data during this period to train the model, the neural network may be more likely to “learn” the differences in the characteristics between the dead cases and the surviving cases and thus improve the prediction power. Additionally, the reverse chronological order approach allowed us to use the latest available clinical data to predict mortality in a real-world setting. Furthermore, using physiological parameters within a longer time to predict mortality may better reflect the reaction of patients to medical treatments. Though it seems not possible to determine a time point that is known in advance as three days before being relieved or dead, our model still has great clinical significance as it could be used for current mortality risk prediction by using patients’ data in the past three days. Thus, it could help physicians identify patients with emergent status by providing dynamic and accurate death risk prediction timely.

Understanding why the deep neural model makes a high death risk prediction for patients can help clinicians make timely interventions, thereby improving clinical outcomes. Attribution analysis based on the integrated gradients method showed that the daily dose of intravenous medications seemed to contribute more substantially to death risk prediction than other features. Patients’ health status are reflected by physiological indicators such as vital signs and laboratory tests results, which would be further understood by physicians and help making diagnosis and treatment. So it is not difficult to understand the use of medications could predict patients clinical outcome. There is no doubt that vasoactive drugs positively affect mortality because they are more often used in patients with shock and hypotension, which are usually associated with a poor prognosis^[26]. Interestingly, we also noticed that the dose of amiodarone, dexmedetomidine and esmolol used on D_3 seemed related to lower death risk, which suggests the early use of these drugs in critically patients when needed. Amiodarone is effective for the

treatment of supraventricular and ventricular arrhythmias, which may partly explain its protective role in these results. However, unwanted extracardiac side effects caused by amiodarone should not be ignored. Parallel to a more frequent use of lower amiodarone, maintenance doses (100–200 mg/day) were recommended by Haverkamp^[27]. Dexmedetomidine has a mild sedative and analgesic effect and is often used for mechanically ventilated patients. It has also been suggested that dexmedetomidine can reduce the prevalence of delirium in critically ill patients and was related to a lower duration of mechanical ventilation and mortality in previous studies^[28, 29]. Esmolol is a beta adrenoceptor antagonist and used primarily for the control of ventricular rate in various emergency status. The result from a randomized clinical trial by Liu^[30] showed that esmolol may decrease the mortality by controlling heart rate in critical patients with septic shock. Although this study attempted to explain the model based on attribution theory, the results demonstrated here are only the tip of the iceberg. Further studies are recommended to report the explainability of deep learning-based predictions to help understand how the models arrive at conclusions at general and individual levels^[31].

Our model also has several advantages over conventional clinical risk tools. First, our model relies entirely on data elements that are likely to be accessible in EHRs, so clinical staff do not have to collect extra variables for a mortality prediction model. It has been estimated that a clinician needs approximately 37 minutes to calculate the APACHE score for one patient^[5]. The cost of manual data collection can thus become a significant barrier to using the tools^[32]. Given the widespread adoption of EHRs, our model has the potential to be integrated into the hospital's intelligent management system and provide real-time and accurate predictions. Second, we incorporated not only data objectively measured during critical care but also pharmacological prescriptions in our model to shape the predictions, reflecting the direct traces of human intelligence. This is a common property of a real clinical setting; however, it has been overlooked in models exclusively trained on systemic patient properties rather than people's reactions to them^[7]. Third, the deep learning method used in this study allows the model to be updated over time. It has been argued that traditional risk tools based on logistic regression methods, such as APACHE scores, should not be used to guide decisions for individual patients because the performance of these tools deteriorates over time, characterized by the worsening of discriminability^[33, 34]. Models based on deep learning technology are not encumbered by such a limitation, however. The key aspect of deep learning is that the features are not designed by human engineers but instead learned from data^[35]. Therefore, the accuracy of the model will likely be improved with larger training sets.

4.2 Limitations

Our study has several limitations. The sample size of the test cohort was relatively small, and the patients were not randomly selected for comparing the prediction results with those of the model based on APACHE II scores, which is one of the most frequently used mortality risk prediction tools in clinical practice. Patients in the test dataset seemed to be more seriously ill because they had a slightly longer length of ICU stay and hospitalization. This may, to some extent, have affected the performance of the validated results. Considering that the test

samples in real-world settings may not always share the same characteristics as the developmental data, the results were acceptable. External validation using data collected from different hospitals or public datasets may further address this limitation. In addition, as the data were imbalanced (with 19.01% mortality prevalence for the whole population), accuracy can be a poor evaluation metric. In addition to the AUROC, we calculated the AUPRC to evaluate the discriminability of the model. Also, the grid search method used for selecting model parameters in our study is somewhat time expensive, more efficient method such as Bayesian-based hyperparameter search is encouraged to be used in further studies. Furthermore, the deep learning model used data collected for up to three days for prediction. Some values were definitively missing for those hospitalized for less than three days and thus need to be carefully processed; otherwise, the model's accuracy may be affected. This problem can be somewhat alleviated by using the clinical experience-based imputation method described above. Further studies are also encouraged to identify whether the deep learning model trained using the same features collected within a shorter period could achieve desirable prediction power in different settings.

5 Conclusions

In conclusion, an accurate model for mortality prediction using data collected from routine workflows is feasible. The novelty of this study is that we developed a model in a real-world setting with better predictive power than the APACHE II scores based on a combination of deep ANNs and the reverse chronological order approach. The explanation analysis suggested that early use of amiodarone and dexmedetomidine when needed may be related to lower mortality. Our results may contribute to developing an intelligent monitoring system for critically ill patients that can monitor their health status and generate timely alerts whenever adverse medical conditions are anticipated. The well-performing and interpretable model may have practical implications for improving the quality of care for critically ill patients.

Acknowledgements This study was partly supported by Xiangya Clinical Big Data Construction Project, a grant from The Central South University. We would like to thank Yang Mingshi, and Yang Bingchang in Critical Care Medicine of The Third Xiangya Hospital of Central South University for their work in medical records review and data check. Also, we would like to thank Liao Hanlong and Sun Yuchen in the College of System Engineering, the National University of Defense Technology, for their work in data preprocessing. We also thank AJE for useful comments and language editing, which have greatly improved the manuscript.

References

- [1] Keegan M T, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. *Critical Care Medicine*, 2011, 39(1): 163–169.
- [2] Salluh J I F, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM. *Current Opinion in Critical Care*, 2014, 20(5): 557–565.
- [3] Knaus W A, Draper E A, Wagner D P, et al. APACHE II: A severity of disease classification system. *Critical Care Medicine*, 1985, 13(10): 818–829.
- [4] Breslow M J, Badawi O. Severity scoring in the critically III: Part 2: Maximizing value from outcome prediction scoring systems. *Chest*, 2012, 141(2): 518–527.

- [5] Kuzniewicz M W, Vasilevskis E E, Lane R, et al. Variation in ICU risk-adjusted mortality: Impact of methods of assessment and potential confounders. *Chest*, 2008, 133(6): 1319–1327.
- [6] Niskanen M, Kari A, Nikki P, et al. Acute physiology and chronic health evaluation (APACHE II) and Glasgow coma scores as predictors of outcome from intensive care after cardiac arrest. *Critical Care Medicine*, 1991, 19(12): 1465–1473.
- [7] Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: A retrospective study. *The Lancet Respiratory Medicine*, 2018, 6(12): 905–914.
- [8] Holmgren G, Andersson P, Jakobsson A, et al. Artificial neural networks improve and simplify intensive care mortality prognostication: A national cohort study of 217,289 first-time intensive care unit admissions. *Journal of Intensive Care*, 2019, 7(1): 1–8.
- [9] Alghatani K, Ammar N, Rezgui A, et al. Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation. *JMIR Medical Informatics*, 2021, 9(5): e21347.
- [10] Reiz A N, de la Hoz M A A, García M S. Big data analysis and machine learning in intensive care units. *Medicina Intensiva*, 2019, 43(7): 416–426.
- [11] Johnson A E W, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016, 3(1): 1–9.
- [12] Pirracchio R, Petersen M L, Carone M, et al. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): A population-based study. *The Lancet Respiratory Medicine*, 2015, 3(1): 42–52.
- [13] Deo R C. Machine learning in medicine. *Circulation*, 2015, 132(20): 1920–1930.
- [14] Shillan D, Sterne J A C, Champneys A, et al. Use of machine learning to analyse routinely collected intensive care unit data: A systematic review. *Critical Care*, 2019, 23: 1–11.
- [15] Petkovic D, Kobzik L, Re C. Machine learning and deep analytics for biocomputing: Call for better explainability. *Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium*, 2018: 623–627.
- [16] Zhao S, Liu P, Tang G, et al. Development and application of an intensive care medical data set for deep learning. *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020: 3369–3373.
- [17] Chadha A, Kaushik B. A hybrid deep learning model using grid search and cross-validation for effective classification and prediction of suicidal ideation from social network data. *New Generation Computing*, 2022, 40(4): 889–914.
- [18] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 2014, 15(1): 1929–1958.
- [19] Boyd K, Eng K H, Page C D. Area under the precision-recall curve: Point estimates and confidence intervals. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23–27, 2013, Proceedings, Part III 13*. Springer Berlin Heidelberg, 2013: 451–466.
- [20] Staffa S J, Zurawski D. Statistical development and validation of clinical prediction models. *Anesthesiology*, 2021, 135(3): 396–405.
- [21] Heller G. The added value of new covariates to the brier score in cox survival models. *Lifetime Data Analysis*, 2021, 27(1): 1–14.
- [22] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. *International Conference on Machine Learning*. PMLR, 2017: 3319–3328.
- [23] Goh G S W, Lapuschkin S, Weber L, et al. Understanding integrated gradients with smoothtaylor for deep neural network attribution. *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021: 4949–4956.
- [24] Delahanty R J, Kaufman D, Jones S S. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. *Critical Care Medicine*, 2018, 46(6): 481–488.
- [25] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436–444.
- [26] Thawitsri T, Chittawatanarat K, Kumwilaisak K, et al. Treatment with vasoactive drugs and outcomes in surgical critically ill patients: The results from the THAI-SICU Study. *Journal of the Medical Association of Thailand*, 2016, 99(Suppl. 6): 83–90.

- [27] Haverkamp W, Israel C, Parwani A. Clinical aspects of treatment with amiodarone. *Herzschrittmachertherapie+Elektrophysiologie*, 2017, 28: 307–316.
- [28] Aso S, Matsui H, Fushimi K, et al. Dexmedetomidine and mortality from sepsis requiring mechanical ventilation: A Japanese nationwide retrospective cohort study. *Journal of Intensive Care Medicine*, 2021, 36(9): 1036–1043.
- [29] Kawazoe Y, Miyamoto K, Morimoto T, et al. Effect of dexmedetomidine on mortality and ventilator-free days in patients requiring mechanical ventilation with sepsis: A randomized clinical trial. *Jama*, 2017, 317(13): 1321–1328.
- [30] Liu H, Ding X F, Zhang S G, et al. Effect of esmolol in septic shock patients with tachycardia: A randomized clinical trial. *National Medical Journal of China*, 2019, 99(17): 1317–1322.
- [31] Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 2020, 20: 1–9.
- [32] Breslow M J, Badawi O. Severity scoring in the critically ill: Part 1-interpretation and accuracy of outcome prediction scoring systems. *Chest*, 2012, 141(1): 245–252.
- [33] Moreno R P, Nassar A P. Is APACHE II a useful tool for clinical research? *Revista Brasileira de Terapia Intensiva*, 2017, 29: 264–267.
- [34] Soares M, Dongelmans D A. Why should we not use APACHE II for performance measurement and benchmarking?. *Revista Brasileira de Terapia Intensiva*, 2017, 29: 268–270.
- [35] Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 2018, 19(6): 1236–1246.