Development of a Machine Learning Model for the Prediction of the Real Time Mortality in Patients in the Intensive Care Unit

Jaeyoung Yang
VHS Medical Center: Seoul Veterans Hospital
Hong-Gook Lim (hongklim@hanmail.net)
Seoul National University Hospital https://orcid.org/0000-0001-9247-6775
Wonhyeong Park
VHS Medical Center: Seoul Veterans Hospital
Dongseok Kim
VHS Medical Center: Seoul Veterans Hospital
Jin Sun Yoon
VHS Medical Center: Seoul Veterans Hospital
Sang-Min Lee
Seoul National University Hospital Department of Internal Medicine
Kwangsoo Kim
Seoul National University Hospital

Research

Keywords: Machine learning, Mortality, Intensive care units, Prognosis, Risk assessment

Posted Date: November 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1066192/v1

Abstract

Background

Prediction of mortality in intensive care units is very important. Thus, various mortality prediction models have been developed for this purpose. However, they do not accurately reflect the changing condition of the patient in real time. The aim of this study was to develop and evaluate a machine learning model that predicts short-term mortality in the intensive care unit using four easy-to-collect vital signs.

Methods

Two independent retrospective observational cohorts were included in this study. The primary training cohort included the data of 1968 patients admitted to the intensive care unit at the Veterans Health Service Medical Center, Seoul, South Korea, from January 2018 to March 2019. The external validation cohort comprised the records of 409 patients admitted to the medical intensive care unit at Seoul National University Hospital, Seoul, South Korea, from January 2019 to December 2019. Datasets of four vital signs (heart rate, systolic blood pressure, diastolic blood pressure, and peripheral capillary oxygen saturation [SpO2]) measured every hour for 10 h were used for the development of the machine learning model. The performances of mortality prediction models generated using five machine learning algorithms, Random Forest (RF), XGboost, perceptron, convolutional neural network, and Long Short-Term Memory, were calculated and compared using area under the receiver operating characteristic curve (AUROC) values and an external validation dataset.

Results

The machine learning model generated using the RF algorithm showed the best performance. Its AUROC was 0.922, which is much better than the 0.8408 of the Acute Physiology and Chronic Health Evaluation II. Thus, to investigate the importance of variables that influence the performance of the machine learning model, machine learning models were generated for each observation time or vital sign using the RF algorithm. The machine learning model developed using SpO2 showed the best performance (AUROC, 0.89).

Conclusions

The mortality prediction model developed in this study using data from only four types of commonly recorded vital signs is simpler than any existing mortality prediction model. This simple yet powerful new mortality prediction model could be useful for early detection of probable mortality and appropriate medical intervention, especially in rapidly deteriorating patients.

Background

Multiple reports have indicated that changes in vital signs often precede the rapid deterioration of a patient's condition [1,2]. This has led to the development of various models for predicting mortality [3,4].

The most well-known mortality prediction models are the Acute physiology and Chronic Health Evaluation (APACHE) system, the Simplified Acute physiology Score (SAPS), and the Mortality Probability Model (MPM) [5-7]. These classic mortality prediction models were developed based on logistic regression, which is easy to interpret. However, although these models show excellent predictive power in actual clinical practice, they have some limitations, such as the inability to handle non-linear relationships and limited handling of internal interactions among predictor variables [8].

In 1950, Alan Turing proposed the Turing test, which allows a human-like response in a computer (also known as an imitation game). However, the concept of machine learning was first introduced by Arthur Samuel in 1959, with his groundbreaking game of computer checkers [9]. Machine learning is a computer algorithm that uses sample data, called training data, to create mathematical models for prediction or decision-making on its own without the need for sophisticated programming. Traditional statistics produce models that focus on the cause and effect between variables in the data, whereas machine learning produces models that focus only on predictive power [9]. Over the past few decades, machine learning has led to remarkable advances in algorithms, and in recent years, its applications in the medical field have been expanding.

Given that the existing mortality prediction models predict mortality using data recorded on the first day of admission to the intensive care unit, they are fundamentally limited in predicting the prognoses of patients with dynamically fluctuating vital signs.

Therefore, the purpose of this study was to develop and evaluate a machine learning-based short-term mortality prediction model for the assessment of real-time mortality in patients in the intensive care unit (ICU).

Methods

Data sources

We used the electronic health record (EHR) data of all patients older than 18 years admitted to the ICU at the Veterans Health Service Medical Center, Seoul, South Korea, from January 2018 to March 2019. The primary training cohort included data from 1968 patients. In cases where a patient had two or more hospitalizations in the ICU, only the data of the last hospitalization were used to prevent duplication of data. A separate ICU dataset containing the records of 409 patients admitted from January 2019 to December 2019 was provided by Seoul National University Hospital, Seoul, South Korea, for external validation. All the data used in this study were anonymized. This study was approved by the institutional review boards of both hospitals (BOHUN 2020-01-031-001, H-2002-063-1100).

Variable selection and data preprocessing

Data on five vital signs (heart rate [HR], systolic blood pressure [SBP], diastolic blood pressure [DBP], respiratory rate [RR], and peripheral capillary oxygen saturation [SpO2]) measured every hour for 10 h

were collected for the development of the machine learning-based model. The vital signs of patients admitted to the ICU were recorded at least once per hour, and the values were automatically stored in the EHR.

Data preprocessing was performed for analysis. All non-numeric data were treated as missing values. If vital sign values exceeded the defined range of physiologically possible values, the patient data were removed and were not used for the development of the model (HR >300 beats per minute, SBP DBP > 300, DBP >200 mmHg, and SpO2>100%). The data of 32 patients were excluded based on this criterion. Thus, the data of 1936 patients were used for the development of the model. If vital signs were measured multiple times within 1 h, the last measurement was used for analysis. The same preprocessing procedure was adapted to the external validation dataset. The record of one patient out of the 409 in the external validation cohort was excluded.

More effort was put into preprocessing the data, especially when dealing with missing values. Approximately 1.8% of the ICU data used for model training had missing values. Although it cannot be said that the proportion of missing values was high compared to the size of the entire dataset, the problem was that the distribution of the missing values was uneven. The ICU data recorded at time points far from the time a patient was discharged from the ICU (e.g., 10 hours before leaving the ICU) often showed omission rates less than 0.5%. However, ICU data recorded just before the patient left the ICU showed omission rates up to 20% in most cases (Figure 1).

All patient data from the training dataset with missing values were removed to generate a complete dataset without missing values (data from 1239 patients in the ICU). Values in each vital sign column were randomly deleted in the same proportions as seen in Figure 1. Missing values in the artificially-generated dataset were filled with a proportion of missing values similar to that of the original data using five well-known missing-value imputation packages in R (multivariate imputation by chained equation [MICE], Amelia, MissForest, Hmisc, and MI). After filling in the missing values using each R package, the performance was assessed using the normalized root mean square error (NRMSE), which is defined as

$$\sqrt{\frac{mean((X^{true} - X^{imp})^2)}{var(X^{true})}}$$

where X^{true} and X^{imp} are derived from the original complete data matrix and the imputed data matrix, respectively, whereas var represents the variance calculated only for consecutive missing values [10]. MissForest showed that the lowest NRMSE values were used for imputation (Figure 2).

Development of the machine learning model

A machine learning method was applied for the prediction of short-term mortality in patients in the ICU. Model development was performed using five machine learning algorithms: Random Forest (RF), XGboost, multi-node perceptrons, convolutional neural network (CNN), and Long Short-term Memory

(LSTM). RR was included as one of the features of the first dataset collected for model development. However, the respiratory rate of a critically ill patient who underwent airway intubation can be adjusted arbitrarily by medical staff; thus, it was excluded from the features used in developing the final model.

Given that the machine learning model has the risk of leaking information about survival and death, the last vital sign recorded upon discharge from the ICU was deleted and not used for machine learning development. Thus, the ICU data of four vital signs recorded over 10 h were used.

Common sense indicates that spending more time data on model development provides better predictive power. However, this only predicts mortality in the near future. Considering that the time taken to manage patients in the ICU appropriately is becoming smaller, the value of the model will decrease substantially. Predicting mortality in the distant future can provide more time for appropriate medical care in the ICU. However, the small amount of available data makes it difficult to build better models. For the development of our model, we considered a balance between predictive power and how well ICU mortality could be predicted in advance. We believed that if the model could predict mortality four hours in advance, it would have both practicality and good predictive ability. Thus, development of a model that could predict mortality four hours in advance was the basic aim of this study.

The area under the receiver operating characteristic curve (AUROC) was primarily used to evaluate the ability to predict survival or death of patients in the ICU. The ICU data were divided into a training dataset (2/3) and a test dataset (1/3). To overcome overfitting and selection bias problems, we subjected the training data to five-fold cross-validation for the optimization of hyperparameters and the selection of models.

For the development of the perceptron model, we used a single-layer feed-forward neural network and tried all combinations of the following settings to find the optimal hyperparameter: the number of first layer nodes (40,42,44,45,46,48,50,52), the number of hidden layer nodes (22,24,26,28,30,32,34), and dropout (0.1,0.2,0.3,0.4). The rectified linear unit (ReLu) was used as the activation function for one input layer and one hidden layer. Sigmoid was used as the activation function for the output layer.

The development of the CNN model consisted of the use of an embedding layer, one-dimensional convolutional operations followed by max pooling, and one neural input layer with the ReLu activation function. All combinations of the following settings were used to find the optimal hyperparameter: the number of dimensions in the embedding layer (32,36,40,44,48,52,56,60,64), the number of neural nodes after max pooling (16,20,24,28,32,36,40,44), and dropout (0.1,0.2,0.3,0.4). Sigmoid was used as the activation function for the output layer.

The development of the LSTM model consisted of the use of an embedding layer and an LSTM layer. All combinations of the following settings were used to find the optimal hyperparameter: the number of dimensions in the embedding layer (32,36,40,44,48,52,56,60,64), the number of LSTM nodes (16,20,24,28,32,36,40,44), and dropout (0.1,0.2,0.3,0.4). Sigmoid was used as the activation function for the output layer.

The default RF package was used in the development of the RF model. All combinations of the following settings were used to find the optimal hyperparameter: the number of trees in the forest (1,2,4,8,16,32,64,100,200) and the minimum number of data points before node split (0.1,0.5,5).

The XGBoost model was developed using default settings without manual parameter adjustments. Default parameter values were used for all other parameters not mentioned. Machine learning was performed using the author's own Keras scripts written in the Python language under Scientific Python development (Spyder) (Keras version: 2.2.5 backend: tensorflow 1.15.0 python 3.7).

Explainable machine learning

The **Sh**apley **A**dditive ex**P**lanation (SHAP) algorithm was used to overcome the lack of explanation for machine learning decisions, which is known as the black box issue. The SHAP explanation method uses coalitional game theory to calculate a Shapley value that represents the extent to which each feature contributes to predicting an outcome. The SHAP value was calculated using the TreeShap algorithm, which can be applied to tree-based machine learning, such as decision trees and RF [11].

Statistical analysis

The performance of the machine learning algorithms (perceptron, XGboost, LSTM, CNN, and RF) and APACHE II scores were compared using AUROC. Accuracy, sensitivity, specificity, positive predicted value, and negative predicted value were calculated for all predictive models in this study. Statistical analyses were performed using Rex (Version 3.0.3, RexSoft Inc., Seoul, Korea) and R 3.5.1 (R Development Core Team; R Foundation for Statistical Computing, Vienna, Austria).

Results

Descriptive statistics

The ICU data of 1968 patients were obtained for the development dataset. The records of 32 patients were excluded because they had implausible data. Thus, the data of 1936 patients were used for the development of the model. The mean age of the patients was 75.38 ± 8.60 years, and 1742 of them were male. A total of 300 patients (15.2 %) died in the ICU (Table 1, Figure 3).

Table 1. Descriptive statistics of the data set

	VHS ICU data	SNU ICU data
Total patient number	1968	409
Age	75.38 ± 8.60	65.63 ± 14.72
Sex (male, %)	1742(88.5%)	409(61.6%)
Mortality	15.2	30.9
Apache II score	19.56 ± 9.79	27.62 ± 11.31
Postoperative patient number	804	0

Selection of the machine learning algorithm

The performance of the models generated using five machine learning algorithms were calculated using AUROC values and the external validation dataset. The performance of the RF model (AUROC, 0.922; 95% confidence interval [CI], 0.881-0.951) was better than that of the other four algorithms (XGboost: 0.903, 0.86-0.933; perceptron: 0.843, 0.789-0.879; CNN: 0.834, 0.793-0.876; LSTM: 0.746, 0.694-0.793) in the four-hour pre-mortality prediction (Figure 4).

In the external validation set, the APACHE II score had an AUROC of 0.84 (95% CI, 0.805-0.879). Thus, we decided to use the RF algorithm for model development because of its superior performance.

Effect of observation time and category selection on model performance

To investigate the importance of variables that influence the performance of the machine learning model, machine learning models were generated for each observation time or vital sign using the RF algorithm.

Datasets of four vital signs recorded over a 10-hour period were the primary data for this study. Using each vital sign data recorded from 5, 6, 7, 8, 9, and 10 h before discharge to the time of discharge, a model for predicting intensive care unit mortality after 6, 5, 4, 3, 2, and 1 h was developed using the RF algorithm.

To determine the best vital signs for predicting mortality, a model was developed using only one of the four vital signs recorded during the seven hours spent on model development. Comparison of the performance of the models based on AUROC showed that the model based on SpO2 had the best performance. The order of the performance of the models were as follows:-SpO2 >DBP >SBP> HR (Figure 5).

Relative importance of each feature to model performance

Comparison of the performance of models using a single variable or multiple observation times has the disadvantage of being unable to investigate complex interrelationships between multiple variables.

A SHAP algorithm was applied to the machine learning model to determine the effect of a single feature on its output in the presence of correlated features. The SHAP values, calculated using the game theoretical approach, can represent importance of features (the magnitude of the contribution) and directions (sign) to the output of the model [12].

The machine learning model generated using the RF algorithm was analyzed using the SHAP algorithm with a 7-hour vital sign dataset. The results indicated that the higher the last recorded SpO2 and SBP levels, the better the patient's chance of survival (Figure 6).

Discussion

In this study, we developed a short-term mortality prediction model for patients in the ICU using four vital signs (HR, SpO2, DBP, and SBP) recorded at one-hour intervals. The performances of mortality prediction models constructed using five machine learning algorithms (RF, XGboost, perceptron, CNN, and LSTM) were evaluated using AUROC values. The best performance was achieved with the RF algorithm (0.922, 0.881-0.951). To avoid the black box problem observed in machine learning models, we applied the SHAP algorithm to our model to explain it.

Models that predict mortality in critically ill patients are already available. At present, the APACHE II, SAPS, and MPM are the most popular models [5-7]. Although the clinical usefulness of these models is already well known, their major drawback is that they use several static physiological parameters to predict patient risk early during hospitalization. Thus, the application of static predictive models to rapidly changing clinical situations in the ICU is limited. Changes in the medical technology environment, such as improvements in computer performance and predictive software algorithms, and expansion of the use of electronic medical records, have made it possible to develop dynamic tools that can predict patient risk in real time.

Thorsen-Meyer reported that real-time LSTM model predicting 90-day mortality was on a 15625 ICU hospitalization dataset [13]. *Meyer* described the development of a real-time recurrent neural network (RNN)-based model for the prediction of postsurgical complications, such as mortality, renal failure, and bleeding, in cardiac patients [14]. *Kim* reported that a real-time mortality prediction model based on a CNN algorithm was created using the datasets of pediatric ICUs [15]. All authors of previous similar reports insisted that their new models outperformed the corresponding clinical reference tools.

Given that it is challenging to evaluate the performance of a model by collecting patient data in real time for research purposes, it is common to retrospectively validate a model using an independent external dataset. However, whether the evaluation is retrospective or prospective, there are always missing values in the data collected. How these missing values are dealt with is always of great interest to researchers. In the present study, we handled missing values in our dataset by deleting the relevant data or replacing the missing values. In most cases, imputation of missing values was performed when there were not too many missing values.

Imputation of missing values is the most important and sophisticated step in data processing for machine learning modeling. In the present study, the missing-value imputation method used in the processing of the training dataset was also used for the validation dataset. It is not easy to recognize the problem of incorrect imputation of missing values until the model is tested on new real data to show its real performance.

In simple datasets, it is possible to replace missing values in some similar summary statistics with the mean or median of the dataset. However, multiple imputation (model imputation) using complex statistics or machine learning models is more popular at present. Multiple imputation is an advanced statistical method of replacing missing values with a number of plausible data values computed using various methods, such as MICE, RF algorithm, and k-nearest neighbor [16]. In this study, we used five imputation models (MICE, Amelia, MissForest, Hmisc, and MI) provided in the R package.

In our internal validation dataset, most of the missing data were found in the datasets of the last three hours (68.3%). It is not clear what causes missing values just before patient discharge. It is possible that the medical staff accidentally entered an incorrect value while preparing the patient for discharge from the ICU, or that vital sign monitoring was turned off while the patient was waiting to be moved to the ward.

In the present study, the basic mortality prediction model did not include vital signs from the last three hours. However, since the trend of vital signs until discharge may affect the quality of the model, we performed missing-value imputation for the entire dataset for 10 hours. The unequal distribution of missing values made it difficult to assess the reliability of the missing-value imputation method. To overcome this drawback, the data were randomly deleted according to time and vital signs at a rate similar to that of the original data after removing all patient data with missing values from the dataset. Thereafter, missing values were inserted using the five multiple imputation methods mentioned above. In the evaluation of imputation methods using NRMSE, Missforest showed the best performance in the dataset used in this study.

Two types of machine learning algorithms were used in this study: neural networks (multilayer perceptron, CNN, and LSTM) and tree types (RF and XGboost). The main difference between the two algorithms is that the decision trees that predict the output label in tree types are independent of each other, whereas the neurons of neural networks are interconnected [17-19].

Three types of neural network architectures were used in this study: multilayer perceptron (MLP), CNN, and LSTM. The MLP is a classical algorithm that consists of three layers (input, hidden, and output). The MLP can process inputs in the forward direction. Any nonlinear function can be optimized using MLP [20]. A CNN consists of three layers: convolution, pooling, and output. Convolution layers are used as filters to extract relevant features from input data. In addition, CNNs show impressive performance in image classification [19]. An LSTM network is a type of recurrent neural network. It specializes in sequential data processing. A block of memory-storing information is attached to the RNN to facilitate the learning of temporal relationships [21].

The basic concept of tree-based algorithms is to create a series of if-then rules to predict the output from the input. There is a large difference between the RF algorithm and the XGboost algorithm in model generation. RF randomly resamples data to build a better model, whereas XGboost improves the model by passing problems found in the previous model to the next model [22]. In this study, RF performed best in predicting ICU mortality. Most of the machine learning algorithms used in this study performed better than the APACHE II score. The APACHE II score is based on initial patient data recorded immediately after admission, whereas the model developed in this study is based on patient data recorded close to the time of discharge, a situation which is not suitable for the APACHE II model. We analyzed the development of the model using selective features and the SHAP algorithm to specifically observe how the model behaves. As a result of examining the performance of the model at various observation points, a better model was developed using the long-term vital sign data recorded close to the time of discharge.

We compared the performance of machine learning models generated using a single vital sign. The model developed with only SpO2 data showed the best performance, whereas the model developed with only HR showed the worst performance. This is somewhat consistent with the results of an analysis conducted using the SHAP algorithm, which showed the strongest positive correlation between the last recorded SpO2 value and the survival of the patient. Further research is needed to determine whether early prediction of mortality using the model developed in this study can improve a patient's prognosis. In addition, considering that the last vital sign data recorded before a patient is discharged were used in the development of this model, further research is needed to determine whether the model can be useful in predicting patient mortality, such as sudden cardiac arrest during hospitalization.

Faced with the challenge of managing the spread of the novel coronavirus disease (COVID-19), the Korean government entrusted our hospital with the operation of living and treatment support centers (LTSCs) for the management of clinically healthy patients with COVID-19. We implemented information and communications technology (ICT)-based remote patient management systems at a COVID-19 LTSC by adopting new electronic health record templates, hospital information system dashboards, cloud-based medical image sharing, a mobile app, and smart vital sign monitoring devices [23]. This mortality prediction model will be useful to optimize outcome by applying our ICT-based tools and applications, which are becoming increasingly important in healthcare.

Conclusions

We developed a model for predicting patient mortality using only four types of commonly recorded vital sign data, which is simpler than any existing mortality prediction model. Further, we showed that using only one of the four vital signs is good for predicting mortality, and can be used in a variety of ways in more restrictive healthcare settings. This simple yet powerful new mortality prediction model could be useful for early detection of probable mortality and appropriate medical intervention, especially in rapidly deteriorating patients.

Abbreviations

APACHE: Acute Physiology and Chronic Health Evaluation; AUROC; Area under the receiver operating characteristic curve; CI: Confidence interval; CNN: Convolutional Neural Network; COVID-19: Novel coronavirus disease; DBP: Diastolic blood pressure; EHR: Electronic health record; HR: Heart rate; ICT: Information and communications technology; ICU: Intensive care unit; LSTM: Long Short-term Memory; LTSCs: Living and treatment support centers; MICE: Multivariate imputation by chained equation; MLP: Multilayer perceptron; MPM: Mortality Probability Model; NRMSE: Normalized root mean square error; ReLu: Rectified linear unit; RF: Random Forest; RNN: Recurrent neural network; RR: Respiratory rate; SAPS: Simplified Acute physiology Score; SBP: Systolic blood pressure; SHAP: Shapley Additive explanation; SpO2: Peripheral capillary oxygen saturation

Declarations

Ethical Approval and Consent to participate

The institutional review boards of Veterans Health Service Medical Center and Seoul National University Hospital granted a waiver of consent for this study (BOHUN 2020-01-031-001, H-2002-063-1100, respectively).

Consent for publication

The authors consent for publication.

Availability of supporting data

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available because they contain information that could compromise research participant privacy.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by Veterans Health Service Medical Center Research Grant, Republic of Korea (VHSMC20034).

Authors' contributions

All authors were involved in the conceptualization and design of the study and the discussion of the results, and approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Acknowledgements

The authors would like to thank Young Lee (Veterans Medical Research Institute, Veterans Health Service Medical Center) for her statistical assistance in preparing the manuscript.

References

1. Berlot G, Pangher A, Petrucci L, Bussani R, Lucangelo U. Anticipating events of in-hospital cardiac arrest. Eur J Emerg Med. 2004;11(1):24-8.

2. Moss TJ, Lake DE, Calland JF, Enfield KB, Delos JB, Fairchild KD, Moorman JR. Signatures of subacute potentially catastrophic illness in the ICU: model development and validation. Crit Care Med. 2016;44(9):1639-48.

3. Clermont G, Angus DC. Severity scoring systems in the modern intensive care unit. Ann Acad Med Singap. 1998;27(3):397-403.

4. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. Ann Intern Med. 1993;118(10):753-61.

5. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818-29.

6. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D. A simplified acute physiology score for ICU patients. Crit Care Med. 1984;12(11):975-7.

7. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA. 1993;270(20):2478-86.

8. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med. 2001;29(2):291-6.

9. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. J Intern Med. 2018;284(6):603-19.

10. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.

11. Lundberg S LS-I. A unified approach to interpreting model predictions. Adv Neur In. 2017;1:4765-74.

12. Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of Computer-Aided Molecular Design. 2020;34(10):1013-26.

13. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, Strøm T, Chmura PJ, Heimann M, Dybdahl L, Spangsege L, Hulsen P, Belling K, Brunak S, Perner A. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. The Lancet Digital Health. 2020;2(4):e179-e91. Epub Mar 12, 2020. doi: 10.1016/s2589-7500(20)30018-2.

14. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, Stamm C, Hofmann T, Falk V, Eickhoff C.. Machine learning for real-time prediction of complications in critical care: a retrospective study. The Lancet Respiratory medicine. 2018;6(12):905-14.

15. Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, Cho I, Park M, Jang H, Kim YH, Kim KW, Sohn MH. A deep learning model for real-time mortality prediction in critically ill children. Critical care (London, England). 2019;23(1):279.

16. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann Transl Med. 2016;4(2):30.

17. Nait Aicha A, Englebienne G, van Schooten KS, Pijnappels M, Krose B. Deep learning to predict falls in older adults based on daily-life trunk accelerometry. Sensors (Basel). 2018;18(5):1654.

18. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

19. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611-29.

20. Kang AR, Lee J, Jung W, Lee M, Park SY, Woo J, Kim SH. Development of a prediction model for hypotension after induction of anesthesia using machine learning. PLoS One. 2020;15(4):e0231172.

21. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80.

22. Davagdorj K, Pham VH, Theera-Umpon N, Ryu KH. XGBoost-based framework for smoking-induced noncommunicable disease prediction. Int J Environ Res Public Health. 2020;17(18):6513.

23. Bae YS, Kim KH, Choi SW, Ko T, Jeong CW, Cho B, Kim MS, Kang E. Information technology-based management of clinically healthy COVID-19 patients: lessons from a living and treatment support center operated by Seoul National University Hospital. J Med Internet Res. 2020;22(6):e19938.

Figures

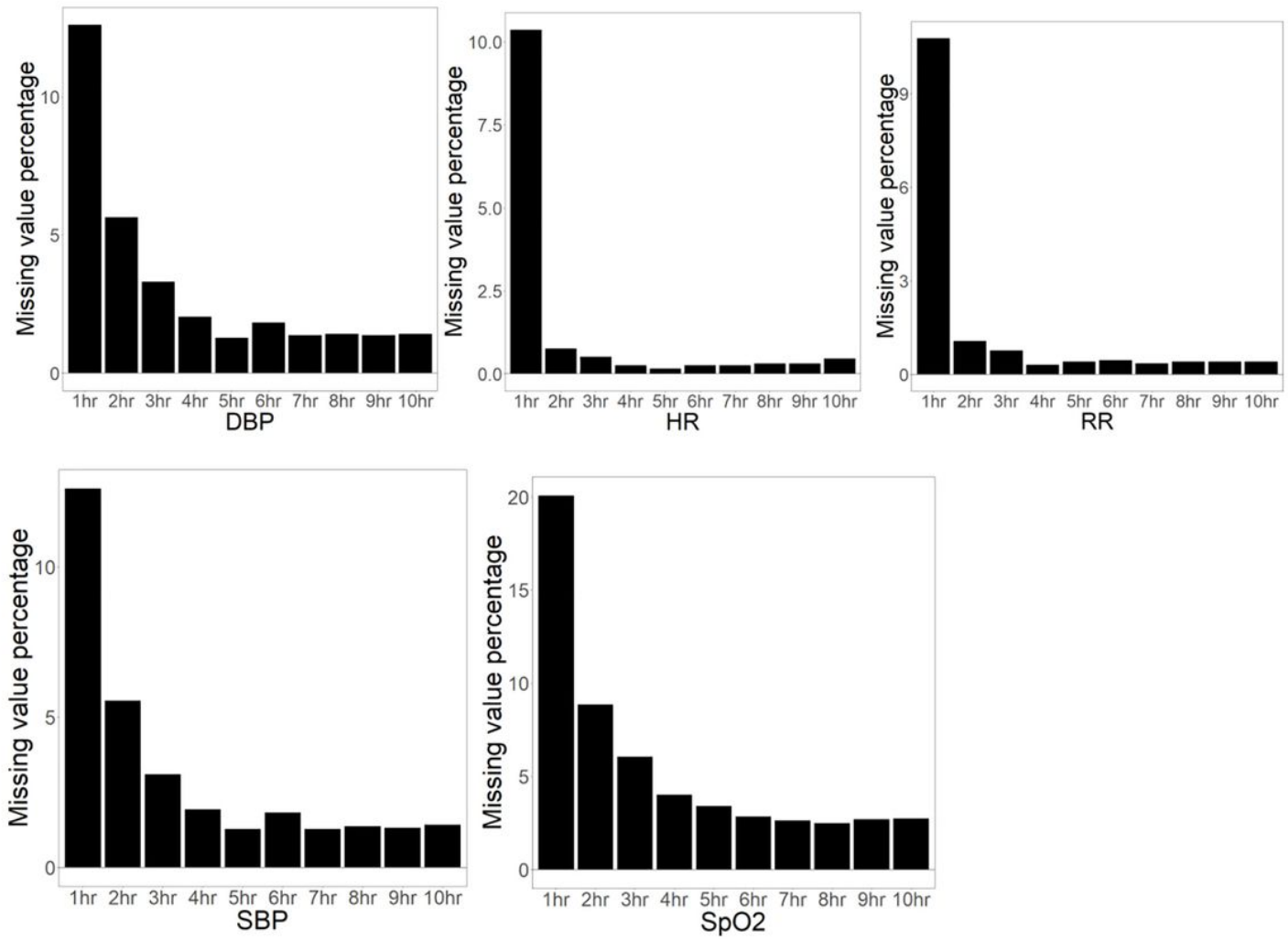


Figure 1

Percentage of missing values for each feature of the five vital signs. The time on the X-axis indicates how many hours before discharge from the intensive care unit. HR, heart rate; SBP, systolic blood pressure; DBP, diastolic blood pressure; RR, respiratory rate; SpO2, peripheral capillary oxygen saturation

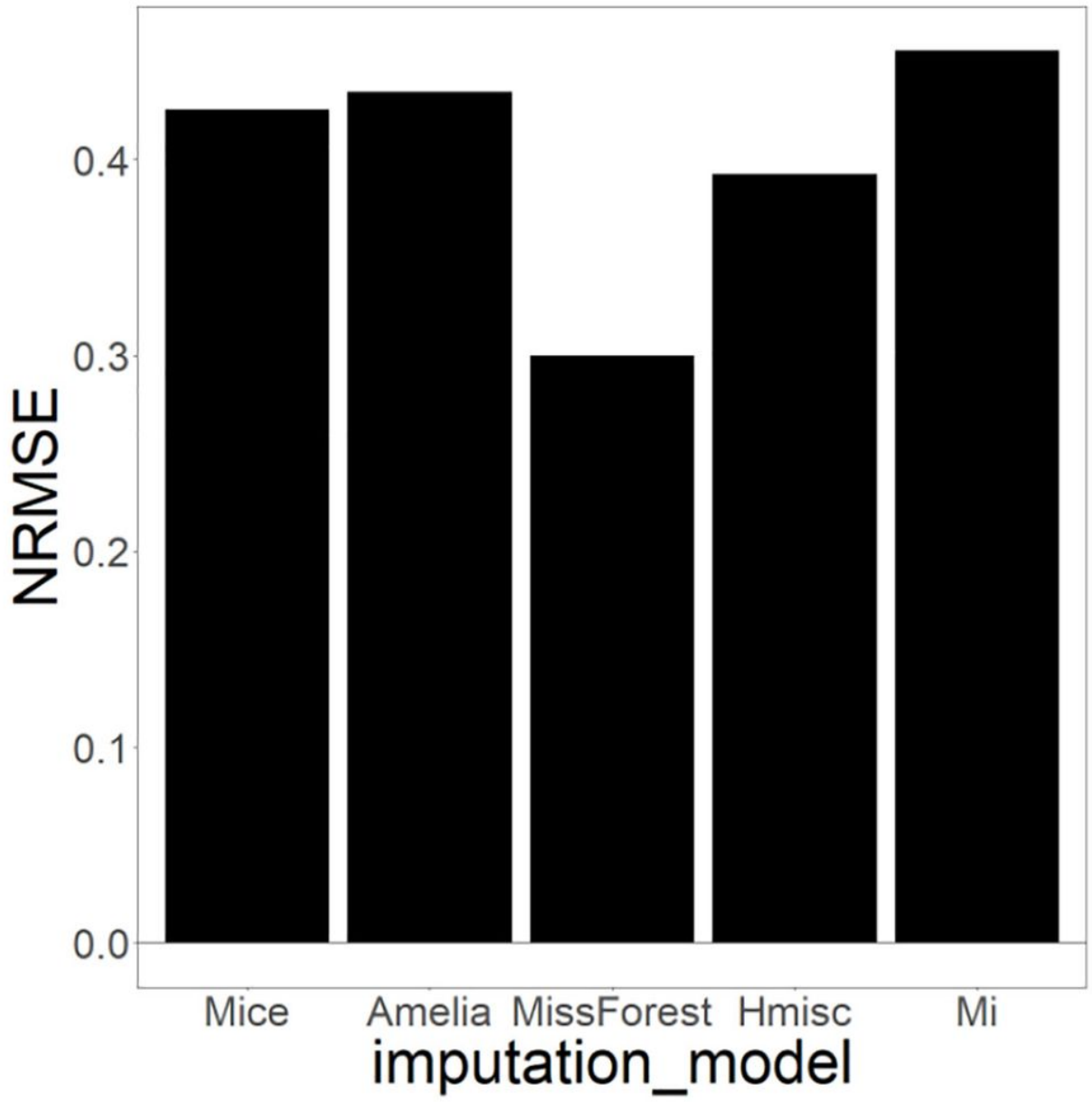


Figure 2

Normalized root mean square error value obtained by imputing missing values of artificially generated datasets using five R packages NRMSE, normalized root mean square error

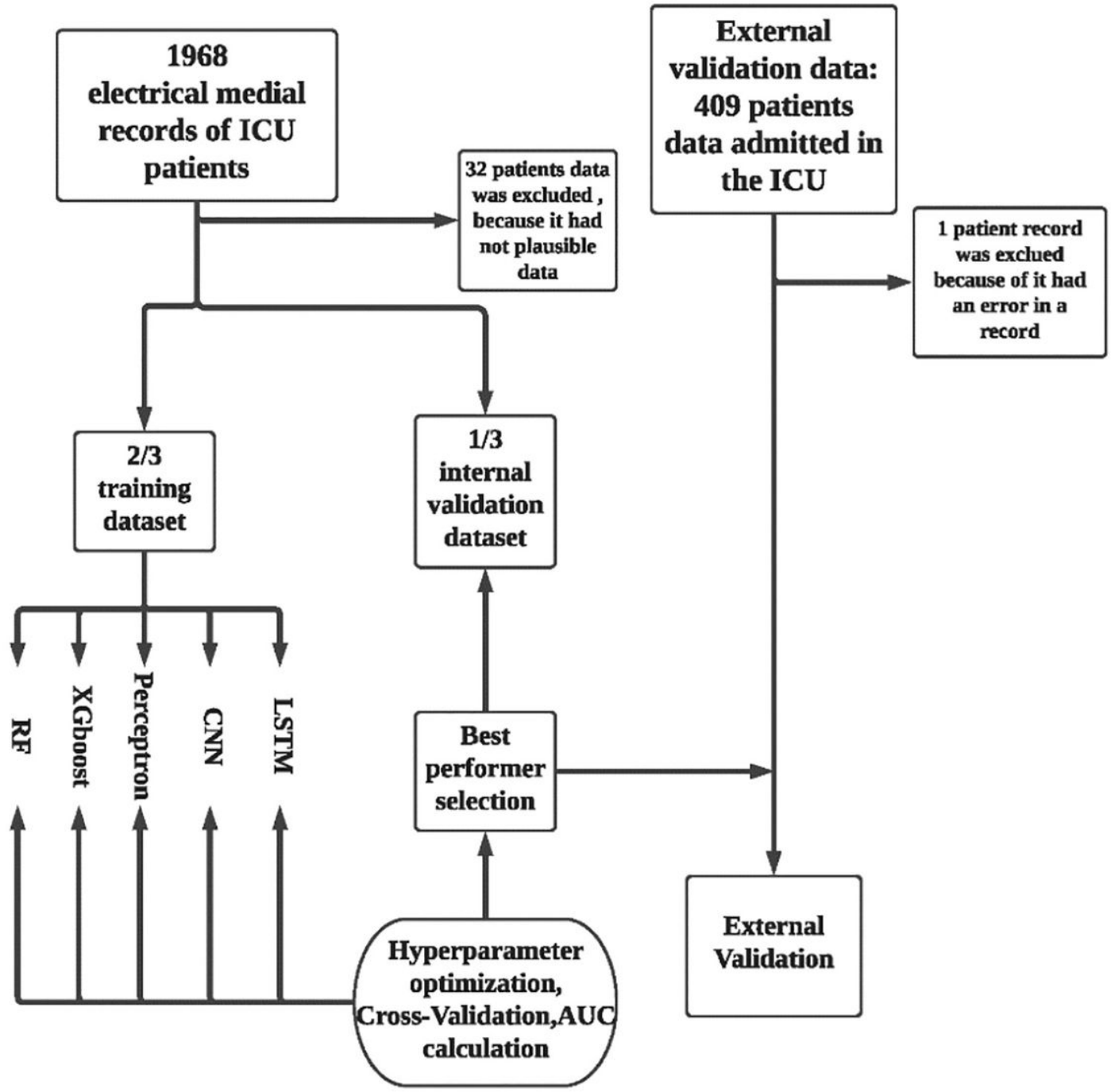


Figure 3

Flowchart of this study ICU, intensive care unit; RF, random forest; CNN, convolutional neural network; LSTM, Long Short-term Memory, AUC, area under the curve

	AUC	sensitivity	specificity	accuracy	ppv	npv
RF	0.922	0.9312	0.8099	0.8942	0.9179	0.8376
Xgboost	0.9039	0.9275	0.7769	0.8816	0.9046	0.8246
perceptrons	0.8433	0.8986	0.6942	0.8363	0.8702	0.75
CNN	0.8358	0.8007	0.7521	0.7859	0.8805	0.6233
LSTM	0.7465	0.8333	0.6116	0.7657	0.8303	0.6167
Apache	0.8408	0.6667	0.8595	0.7254	0.9154	0.5306

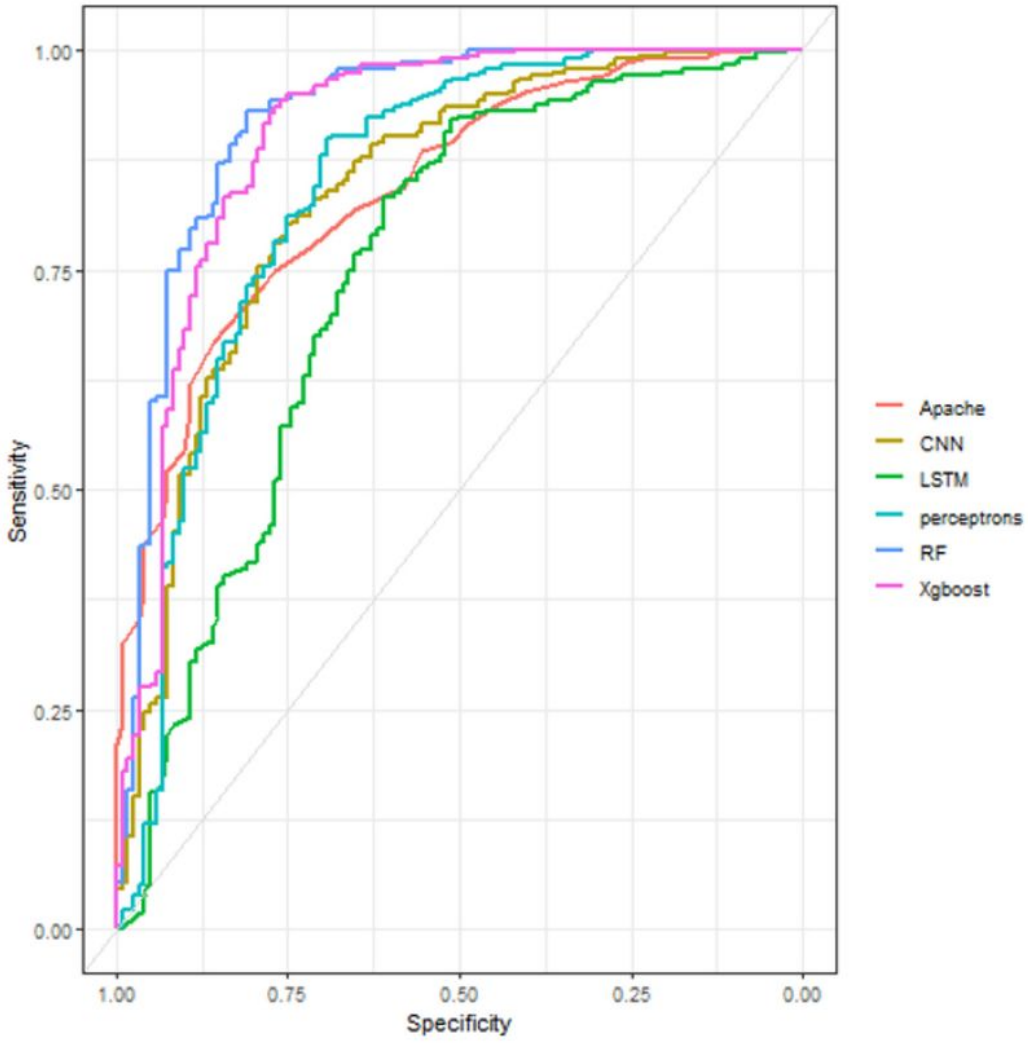
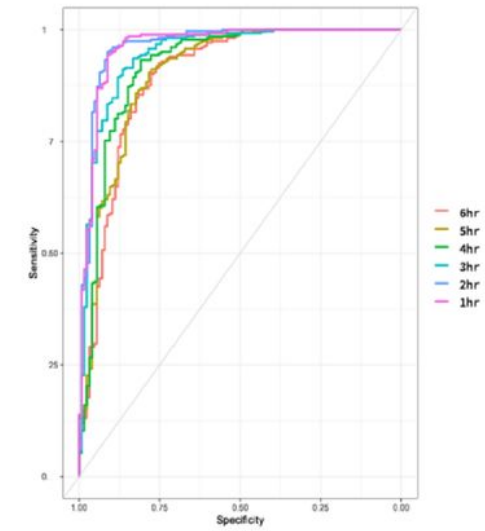


Figure 4

Table(a) and Receiver characteristic operating curve(b) for model performance of five machine learning algorithms for predicting mortality four hours in advance. RF, random forest; CNN, convolutional neural network; LSTM, Long Short-term Memory, APACHE, Acute Physiology and Chronic Health Evaluation

Mortality prediction after N hours	AUC	sensitivity	specificity	accuracy	ppv	npv
6	0.8949	0.9255	0.754	0.8725	0.8938	0.819
5	0.9034	0.9043	0.7857	0.8676	0.9043	0.7857
4	0.9191	0.9326	0.8095	0.8946	0.9164	0.843
3	0.9463	0.9113	0.8651	0.8971	0.938	0.8134
2	0.964	0.9504	0.9127	0.9387	0.9606	0.8915
1	0.9632	0.9433	0.9127	0.9338	0.9603	0.8779



Single variable for model development	AUC	sensitivity	specificity	accuracy	ppv	npv
DBP	0.863	0.8262	0.8016	0.8186	0.9031	0.6733
HR	0.7512	0.5957	0.8016	0.6593	0.8705	0.4698
SBP	0.8627	0.8156	0.8016	0.8113	0.902	0.6601
Spo2	0.89	0.9362	0.7937	0.8922	0.9103	0.8475

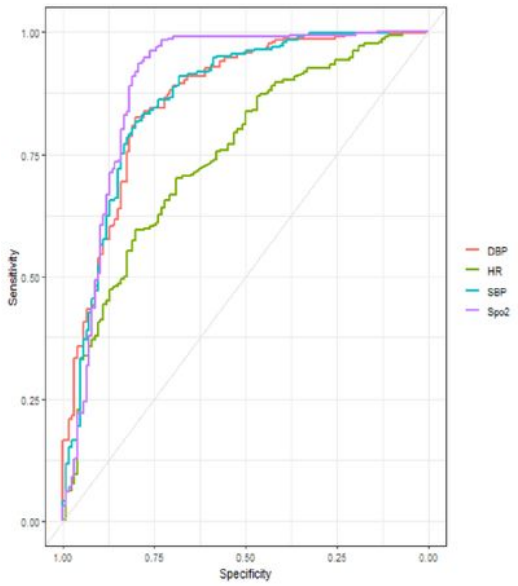


Figure 5

Receiver operating characteristic curves and tables for model performance according to observation time (a,b) and category selection (c,d) for predicting mortality four hours in advance. HR, heart rate; SBP, systolic blood pressure; DBP, diastolic blood pressure; SpO2, peripheral capillary oxygen saturation

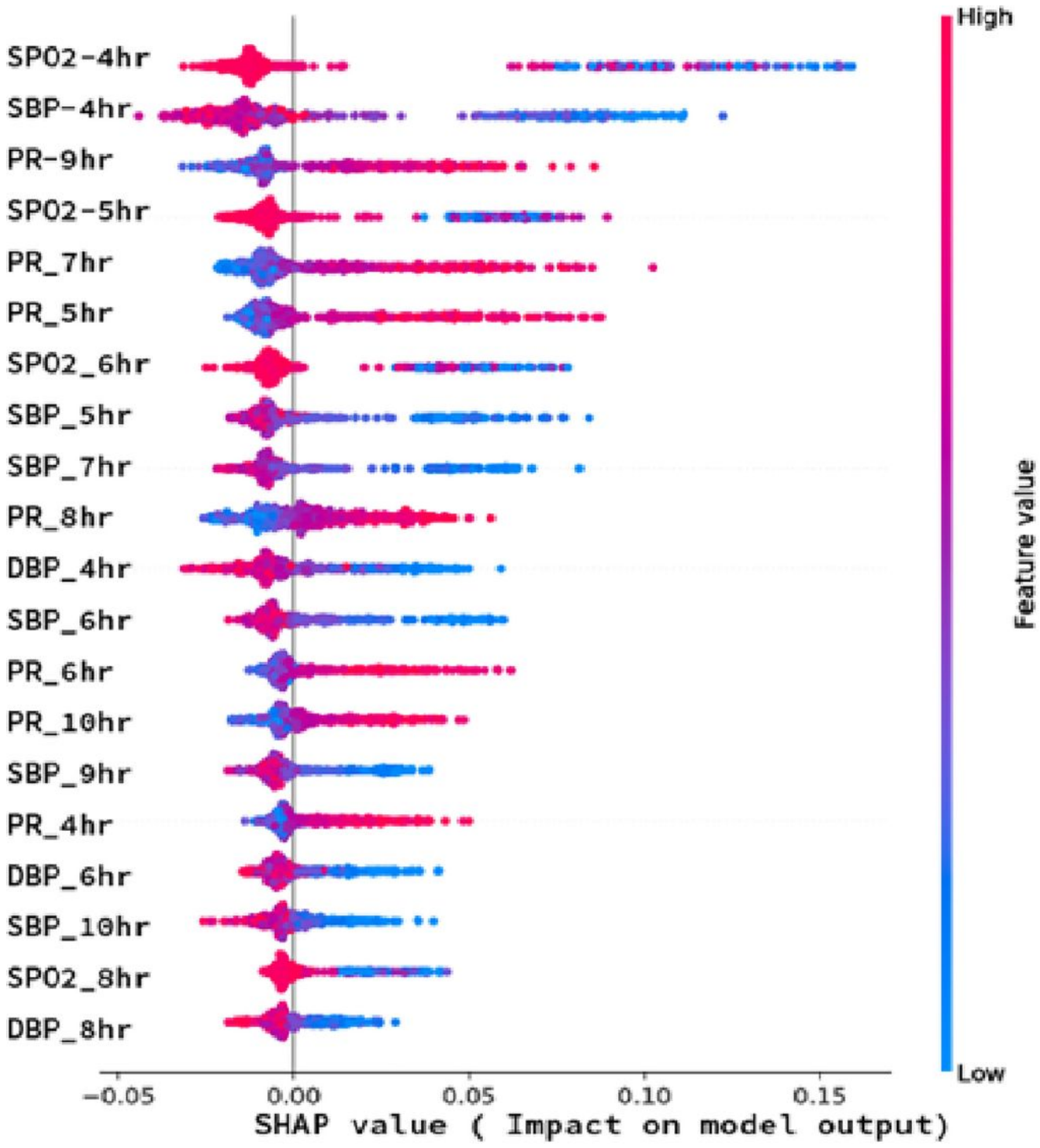


Figure 6

Shapley Additive exPlanation value graph for the random forest machine learning model. The SHAP values of the 20 most important features were displayed in descending order with SHAP values. In X-axis, positive SHAP value indicates more influence on patient death. HR, heart rate; SBP, systolic blood pressure; DBP, diastolic blood pressure; SpO2, peripheral capillary oxygen saturation