Development and validation of a machine-learning model for prediction of hypoxemia after extubation in intensive care units

Ming Xia[#], Chenyu Jin[#], Shuang Cao, Bei Pei, Jie Wang, Tianyi Xu, Hong Jiang

Department of Anesthesiology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Contributions: (I) Conception and design: H Jiang, M Xia, C Jin; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: C Jin; (V) Data analysis and interpretation: M Xia, C Jin; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work and should be considered as co-first authors.

Correspondence to: Hong Jiang. Department of Anesthesiology, Shanghai Ninth People's Hospital, 639 Zhizaoju Road, Shanghai, China. Email: jianghongjiuyuan@163.com.

Background: Extubation is the process of removing tracheal tubes so that patients maintain oxygenation while they start to breathe spontaneously. However, hypoxemia after extubation is an important issue for critical care doctors and is associated with patients' oxygenation, circulation, recovery, and incidence of postoperative complications. Accuracy and specificity of most related conventional models remain unsatisfactory. We conducted a predictive analysis based on a supervised machine-learning algorithm for the precise prediction of hypoxemia after extubation in intensive care units (ICUs).

Methods: Data were extracted from the Medical Information Mart for Intensive Care (MIMIC)-IV database for patients over age 18 who underwent mechanical ventilation in the ICU. The primary outcome was hypoxemia after extubation, and it was defined as a partial pressure of oxygen <60 mmHg after extubation. Variables and individuals with missing values greater than 20% were excluded, and the remaining missing values were filled in using multiple imputation. The dataset was split into a training set (80%) and final test set (20%). All related clinical and laboratory variables were extracted, and logistics stepwise regression was performed to screen out the key features. Six different advanced machine-learning models, including logistics regression (LOG), random forest (RF), K-nearest neighbors (KNN), support-vector machine (SVM), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM), were introduced for modelling. The best performance model in the first cross-validated dataset was further fine-tuned, and the final performance was assessed using the final test set.

Results: A total of 14,777 patients were included in the study, and 1,864 of the patients' experienced hypoxemia after extubation. After training, the RF and LightGBM models were the strongest initial performers, and the area under the curve (AUC) using RF was 0.780 [95% confidence interval (CI), 0.755–0.805] and using LightGBM was 0.779 (95% CI, 0.752–0.806). The final AUC using RF was 0.792 (95% CI, 0.771–0.814) and using LightGBM was 0.792 (95% CI, 0.770–0.815).

Conclusions: Our machine learning models have considerable potential for predicting hypoxemia after extubation, which help to reduce ICU morbidity and mortality.

Keywords: Extubation; hypoxemia; machine learning; anesthesiology

Submitted Apr 01, 2022. Accepted for publication May 20, 2022.
doi: 10.21037/atm-22-2118
View this article at: https://dx.doi.org/10.21037/atm-22-2118

Introduction

Many patients in intensive care units (ICUs) need mechanical ventilation for various reasons, including respiratory failure, coma, and postoperative airway management. Patients are extubated when their respiratory functions improve or airway risks are decreased. Extubation is the process of removing tracheal tubes so that patients maintain oxygenation while breathing spontaneously. However, hypoxemia after extubation is an important issue for critical care doctors. Although senior clinicians can make empirical predictions, hypoxemia after extubation is still inevitable and has a serious impact on patients' oxygenation, circulation (1), recovery (2), and incidence of postoperative complications (3,4). Extubation in the ICU is associated with higher risks than extubation in the postanesthesia care unit (PACU). A clinician needs to balance the risks of extubation in the ICU against the risks of delaying extubation in a patient who requires it. At present, studies have explored prediction models and risk factor analysis of hypoxemia after extubation through various methods (5,6). However, because the number of patients included has been limited by objective factors, most of the studies related with hypoxemia after extubation had a small sample size. In the cases of few training samples (7,8), machine learning models generally cannot achieve good out-of-sample performance, and models trained with small samples are prone to overfitting to small samples and underfitting to the target task.

Databases such as the Medical Information Mart for Intensive Care (MIMIC) have been used to build models to predict mortality (9,10) and morbidity (11,12). A predictive model may provide an early warning to clinicians before the manifestation of clinical signs. By collecting and analyzing the clinical data of patients who have undergone mechanical ventilation in the intensive care unit through the MIMIC-IV database, a more accurate and specific prediction model for extubation can be established.

Machine-learning (ML) models based on mathematical and statistical methods can be used to analyze and infer relationships between clinical variables and patient outcomes (13), and they are the core and foundation of artificial intelligence. Machine learning algorithms have some inherent advantages over other conventional algorithms (14). While conventional algorithms require the a priori selection of a model based on the available data, ML allows greater flexibility in model fitting (15). Furthermore, the variables included in traditional algorithms are limited by the sample size. Instead, by design, ML models are able to consider multiple variables at the same time, and as such, have the potential to detect underlying patterns that may otherwise be undetectable when data are examined effectively in individual silos. With the assistance of ML, more precise models can be used for clinical prediction, diagnosis, and decision-making.

The objective of this study was to develop a prediction model utilizing bedside clinical and laboratory parameters by machine learning to predict hypoxemia after extubation in the ICU. This will help ICU clinicians predict the risk of hypoxemia after extubation, thereby helping to reduce ICU morbidity and mortality. We present the following article in accordance with the TRIPOD reporting checklist (available at https://atm.amegroups.com/article/view/10.21037/atm-22-2118/rc).

Methods

Data collection

The present study used data accessed from the MIMIC-IV database (16), which is a publicly available database that contains real hospital stay data for patients admitted to a tertiary academic medical center in Boston, USA between 2008 and 2019. A total of 524,520 medical records are available in the database. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). One author (CY J) obtained access to the database and was responsible for data extraction.

The present study was based on a cohort generated from the existing database. The following inclusion criteria were applied: (I) patients aged above 18 years, and (II) patients who had undergone mechanical ventilation in the ICU. If patients underwent multiple intubations and mechanical ventilation, we only used data from the first mechanical ventilation. Data with low quality, such as cases with missing values greater than 20%, were excluded.

Related clinical and laboratory variables were extracted from the MIMIC-IV database, including baseline patient characteristics, vital signs, the results of laboratory examinations, and mechanical ventilation parameters. Comorbidities were assessed based on the International Classification of Disease (ICD) codes ICD-9-clinical modification (CM) and ICD-10 (17). Some repeatedly recorded variables were extracted as the maximum, minimum, and final values (the final value was defined as the final recorded data before extubation). Urine output and Sequential Organ Failure Assessment (SOFA) scores were recorded and extracted 24 hours before extubation. The time window for extracting the clinical and laboratory

variables was from ICU admission to extubation. All variables are shown in Table S1.

To include as much data as possible, for the values that were missing and excluded from the analysis, we estimated the relationship between the feature numbers and missing data threshold and yielded 80% as the threshold, which was consistent with the 1:20 principle to avoid overfitting (18). The primary outcome was hypoxemia after extubation, and it was defined as a partial pressure of oxygen (PaO_2) <60 mmHg after extubation.

Multivariate imputation

Multivariate imputation was conducted through an iterative imputer using the R package Multivariate Imputation by Chained Equations (MICE). The multivariate imputation procedure can be split into following steps (19): Step 1: a simple imputation is performed for each missing value in the dataset as "place holders"; Step 2: the mean imputations of "place holder" for variable ("var") are inserted back to missing value; Step 3: the values from "var" are regressed on the other variables in the imputation model; Step 4: the missing data for "var" is altered by predictions according to the regression model; Step 5: repeat steps 2–4.

Model selection

Baseline characteristics were compared between the nonhypoxemia group and the hypoxemia group. Six different advanced machine-learning models were introduced, including K-nearest neighbors (KNN), support-vector machine (SVM), logistic regression (LOG), random forest (RF), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) for the modelling. The dataset was first randomly split into a training set (80%) and a testing set (20%). Logistics stepwise regression with the forward method was performed to screen out the key features. The features in the final stepwise model of each 5 multivariate impute databases were screened, and the features included in all 5 screen results were selected for further study. Furthermore, we calculated the threshold-dependent measures of the sensitivity, specificity, and accuracy at the "best" thresholds for all the models. The "best" threshold was the threshold that maximizes both sensitivity and specificity. A 5-fold cross-validation in the 80% training set was conducted in order to reduce the bias caused by the randomly splitted dataset. The model on each dataset was trained and evaluated, and the area under the curve (AUC) was calculated.

Data expansion

We corrected for the bias in the number of cases between the 2 groups by performing data expansion. The data used for training were matched with 6,000 cases (3,000 positive and 3,000 negative cases). For the test data, we determined whether they would be used for the validation set or final test set, and none of the data were expanded. Data expansion was performed using the "ROSE" package in R software.

Parameter tuning

All the models were simply tuned with a small range grid search according to the package default. Parameter tuning refers to optimizing the algorithm for optimal performance by modifying parameters. The best models were tuned for the parameters specific to the method, since modifiable parameters were different for each machine learning algorithm. Tuning parameters were evaluated by extended manual grid search or using functions in the R package, where each tuning parameter gave a large but realistic range of values. The variable importance of the final optimal model was determined by a ML algorithm that was amenable to computing this value. The package used for each ML model and the tuning parameters for each model are shown in Table S2.

Sensitivity analyses

Different definition of hypoxemia after extubation

The definition of severe hypoxemia after extubation was PaO_2 <30 mmHg, which is the value that is associated with more serious complications. We conducted sensitivity analysis in which PaO_2 <30 mmHg was considered severe hypoxemia after extubation and trained the best performing algorithm on this new definition to generate a new model.

Dataset without multiple imputation

Since multiple imputation is based on the assumption of random missing, it is often impossible to verify whether the assumption of random missing is correct in practical applications. Therefore, a sensitivity analysis method is needed to verify the reliability of the results of multiple imputation analysis under the assumption of missing random. We conducted sensitivity analysis in which the missing data was not filling by multiple imputation

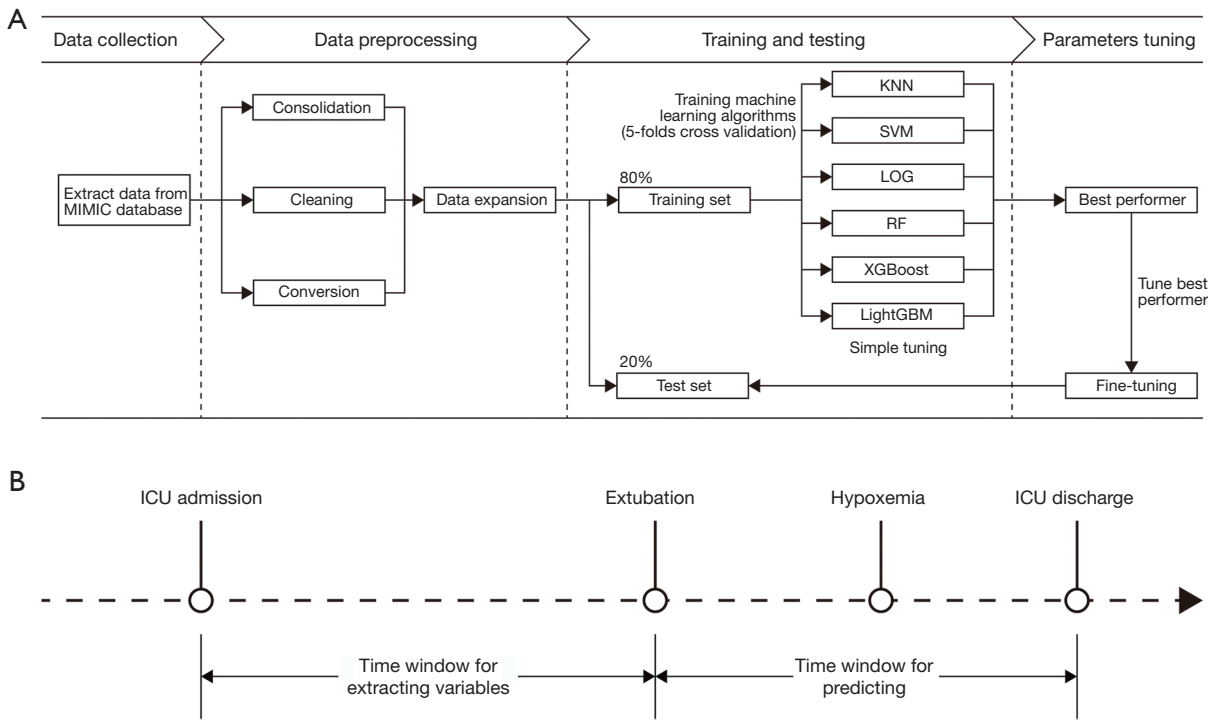

Figure 1 Flow diagram of the study. (A) The study process; (B) the time window for extracting the variables and the predictions. KNN, K-nearest neighbors; SVM, support-vector machine; LOG, logistic regression; RF, random forest; XGBoost, eXtreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; ICU, intensive care unit.

and trained the best performing algorithm on this new definition to generate a new model.

Statistical analysis

The merging and screening of the initial data were performed by Stata (Stata/MP 16.0 for Windows, StataCorp LLC, College Station, TX, USA). Continuous variables with a normal distribution are reported as the mean ± standard deviation. Nonnormally distributed continuous variables are reported as medians (interquartile ranges). Categorical variables are reported as frequencies (percentages). The hypothesis was tested using one-way analysis of variance (ANOVA), the Mann-Whitney U test, and Fisher's exact probability method. Stepwise logistic models were constructed with R. The median of the AUCs was used to evaluate the effectiveness of the model, and the receiver operating characteristic (ROC) curve was shown as the result for each model. An AUC between 0.6–0.7, 0.7–0.8, 0.8–0.9, and 0.9–1.0, was considered to have poor, acceptable, good, and excellent discrimination performance, respectively. DeLong test was used to calculate statistical differences in AUC of different models under the same test set. $P<0.05$ was considered statistically significant. Multiple imputation was performed using the "mice" package in R. ROC curves were drawn using the pROC package in R 4.0.4. The confidence interval (CI) of the AUC was obtained by applying the bootstrap method.

Results

Baseline patient characteristics and variable details

After excluding data with low quality, data with over 20% missing values, and nonfirst-time mechanical ventilation data, 14,777 patients remained, 1,852 (12.5%) of whom experienced hypoxemia after extubation. Ultimately, the training set contained 11,749 cases, and the test set contained 3,028 cases. There were 1,476 (12.6%) cases of hypoxemia after extubation within the training set, and there were 376 (12.4%) cases of hypoxemia after extubation within the test set. The study process is shown in *Figure 1*. Baseline patient characteristics and variable details are shown in *Tables 1*,*2*, respectively.

Table 1 Baseline patient characteristics

Characteristics	Nonhypoxemia	Hypoxemia	P
Number	12,925	1,852	
Age (years)	65.13±14.88	66.44±15.02	<0.001
Gender			<0.001
Male	4,609 (35.7)	816 (44.1)	
Female	8,316 (64.3)	1,036 (55.9)	
Weight (kg)	83.06±21.82	82.68±26.11	0.492
Height (cm)	169.95±11.64	168.27±11.70	<0.001
Coronary heart disease			<0.001
No	6,865 (53.1)	1,160 (62.6)	
Yes	6,060 (46.9)	692 (37.4)	
Hypertension			<0.001
No	6,557 (50.7)	1,136 (61.3)	
Yes	6,368 (49.3)	716 (38.7)	
Pneumonia			<0.001
No	10,610 (82.1)	1,049 (56.6)	
Yes	2,315 (17.9)	803 (43.4)	
Respiratory failure			<0.001
No	8,989 (69.5)	633 (34.2)	
Yes	3,936 (30.5)	1,219 (65.8)	
Diabetes mellitus			<0.001
No	8,864 (68.6)	1,193 (64.4)	
Yes	4,061 (31.4)	659 (35.6)	
Heart failure			<0.001
No	9,719 (75.2)	1,078 (58.2)	
Yes	3,206 (24.8)	774 (41.8)	
Cerebrovascular disease			0.823
No	11,270 (87.2)	1,619 (87.4)	
Yes	1,655 (12.8)	233 (12.6)	
Renal disease			<0.001
No	10,680 (82.6)	1,422 (76.8)	
Yes	2245 (17.4)	430 (23.2)	
Liver disease			<0.001
No	11,653 (90.2)	1,587 (85.7)	
Yes	1,272 (9.8)	265 (14.3)	
Cancer			0.001
No	11,723 (90.7)	1,625 (87.7)	
Yes	1,202 (9.3)	227 (12.3)	

Data are shown as mean ± standard deviation or number (%).

Area under the curve

After training, the AUC using LOG was 0.776 (95% CI, 0.750–0.803); using SVM, it was 0.737 (95% CI, 0.709–0.766); using KNN, it was 0.765 (95% CI, 0.739–0.791); using RF, it was 0.780 (95% CI, 0.755–0.805); using XGBoost, it was 0.704 (95% CI, 0.676–0.732); and using LightGBM, it was 0.779 (95% CI, 0.752–0.806). The ROC, sensitivity, specificity, and accuracy at the best thresholds for each machine-learning method are displayed in *Table 3* and *Figure 2*. The final feature selection after recursive feature elimination is shown in *Figure 3*.

Based on the model selection process, it appeared that the RF and LightGBM models were the strongest initial performers to be candidates for continued tuning and further testing. The other parameters that were tuned specific to the RF and LightGBM methods are shown in Table S2. The final AUC using RF was 0.792 (95% CI, 0.771–0.814) and using LightGBM was 0.792 (95% CI, 0.770–0.815). The final variable importance is shown in *Figure 4*. The specificity was 0.672 (95% CI, 0.584–0.734) in the LightGBM model and 0.669 (95% CI, 0.584–0.737) in the RF model. The sensitivity was 0.801 (95% CI, 0.718–0.883) in the LightGBM model and 0.814 (95% CI, 0.737–0.888) in the RF model. The accuracy was 0.687 (95% CI, 0.618–0.736) in the LightGBM model and 0.686 (95% CI, 0.621–0.734) in the RF model. The ROC, sensitivity, specificity, and accuracy at the best thresholds for each machine-learning method are shown in *Table 4* and *Figure 4*.

The best final AUC using RF and LightGBM was both 0.792. For the final AUC using RF, there was no statistical difference when compared to the AUC using LightGBM (P=0.725), but there were statistical differences when compared to the AUC using KNN (P<0.001), LOG (P=0.033), SVM (P<0.001) and XGBoost (P<0.001). For the final AUC using LightGBM, there was no statistical difference when compared to the AUC using RF (P=0.725) and LOG (P=0.505), but there were statistical differences when compared to the AUC using KNN (P=0.07), SVM (P<0.001) and XGBoost (P<0.001).

Sensitivity analyses

Different definition of hypoxemia after extubation

The AUC using LOG was 0.778 (95% CI, 0.748–0.808); using SVM, it was 0.729 (95% CI, 0.692–0.764); using KNN, it was 0.760 (95% CI, 0.728–0.793); using RF, it was 0.780 (95% CI, 0.748–0.812); using XGBoost, it was 0.707 (95% CI, 0.672–0.741); and using LightGBM, it was 0.777

Table 2 Details of the variables used in the model

Variables	Nonhypoxemia	Hypoxemia	P
Number	12,925	1,852	
Gender			<0.001
Male	4,609 (35.7)	816 (44.1)	
Female	8,316 (64.3)	1,036 (55.9)	
Heart failure			<0.001
No	9,719 (75.2)	1,078 (58.2)	
Yes	3,206 (24.8)	774 (41.8)	
Pneumonia			<0.001
No	10,610 (82.1)	1,049 (56.6)	
Yes	2,315 (17.9)	803 (43.4)	
Respiratory failure			<0.001
No	8,989 (69.5)	633 (34.2)	
Yes	3,936 (30.5)	1,219 (65.8)	
SpO_2 (final)	97.16±5.53	97.02±2.97	0.293
SpO_2 (min)	92.32±9.34	89.68±8.90	<0.001
Respiratory rate (final) (min^{-1})	19.24±5.62	20.43±5.90	<0.001
Respiratory rate (max) (min^{-1})	27.69±8.13	31.95±8.78	<0.001
Heart rate (final) (bpm)	85.46±16.77	88.34±17.23	<0.001
Heart rate (min) (bpm)	66.53±12.98	65.06±14.10	<0.001
RBC (max) (k/μL)	3.73±0.60	3.73±0.69	0.934
RBC (min) (k/μL)	3.10±0.65	2.95±0.68	<0.001
WBC (min) (k/μL)	9.08±5.14	9.14±6.28	0.669
Blood glucose (final) (mg/dL)	134.34±48.64	142.70±54.81	<0.001
Blood glucose (max) (mg/dL)	182.32±112.66	206.59±116.19	<0.001
Lactate (final) (mmol/L)	1.86±1.42	1.57±0.89	<0.001
Lactate (max) (mmol/L)	3.10±2.30	3.43±2.69	<0.001
pH (final)	7.39±0.06	7.40±0.06	<0.001
PaO_2 (final) (mmHg)	121.83±50.92	94.78±48.87	<0.001
PaO_2 (max) (mmHg)	316.02±132.24	244.48±135.95	<0.001
PaO_2 (min) (mmHg)	94.39±50.77	64.79±42.45	<0.001
$PaCO_2$ (final) (mmHg)	40.53±7.20	43.17±10.05	<0.001
Airway pressure (min) (cmH_2O)	6.06±2.95	5.00±2.88	<0.001
PEEP (final) (cmH_2O)	4.91±2.11	4.58±1.92	<0.001
PSV level (final) (cmH_2O)	5.57±2.20	5.63±1.98	0.333
Ventilation time (h)	54.79±82.28	89.15±94.60	<0.001
SOFA (24 h)	5.19±2.97	5.95±3.14	<0.001
SOFA CNS (24 h)	0.66±1.16	0.60±1.00	0.035
Vasopressor			0.003
No	12,868 (99.6)	1,833 (99.0)	
Yes	57 (0.4)	19 (1.0)	

Data are shown as mean ± SD or number (%). RBC, red blood cell; WBC, white blood cell; PEEP, positive end expiratory pressure; PSV, pressure support ventilation; SOFA, Sequential Organ Failure Assessment; SD, standard deviation.

(95% CI, 0.745–0.808). The ROC, sensitivity, specificity, and accuracy at the best thresholds for each machine-learning method are displayed in *Table 5*.

Dataset without multiple imputation

The AUC using LOG was 0.742 (95% CI, 0.707–0.777); using SVM, it was 0.693 (95% CI, 0.655–0.731); using KNN, it was 0.717 (95% CI, 0.679–0.754); using RF, it was 0.751 (95% CI, 0.716–0.787); using XGBoost, it was 0.683 (95% CI, 0.647–0.719); and using LightGBM, it was 0.743 (95% CI, 0.709–0.778). The ROC, sensitivity, specificity, and accuracy at the best thresholds for each machine-learning method are displayed in *Table 6*.

Discussion

In this study, we examined the use of machine-learning methods based on data from the MIMIC-IV database for postoperative predictive analytics, specifically, the prediction of hypoxemia after extubation. The best models that demonstrated better discrimination were the RF and LightGBM models. The AUC using RF was 0.780 (95% CI, 0.755–0.805) in the training set and 0.792 (95% CI, 0.771–0.814) in the test set. The AUC using LightGBM was 0.779 (95% CI, 0.752–0.806) in the training set and 0.792 (95% CI, 0.770–0.815) in the test set. This study developed a prediction model utilizing bedside clinical and laboratory parameters by machine learning to predict hypoxemia after extubation in the ICU.

Many machine-learning algorithms have been utilized

Table 3 ROC, sensitivity, specificity, and accuracy at the best thresholds in the K-fold set

Variables	AUC (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Accuracy (95% CI)
RF	0.780 (0.755–0.805)	0.627 (0.554–0.702)	0.821 (0.731–0.891)	0.653 (0.596–0.710)
KNN	0.765 (0.739–0.791)	0.641 (0.565–0.684)	0.792 (0.728–0.862)	0.661 (0.602–0.694)
LOG	0.776 (0.750–0.803)	0.589 (0.536–0.780)	0.848 (0.647–0.907)	0.621 (0.578–0.767)
SVM	0.737 (0.709–0.766)	0.648 (0.536–0.758)	0.745 (0.614–0.853)	0.659 (0.570–0.743)
XGB	0.704 (0.676–0.732)	0.716 (0.697–0.736)	0.691 (0.638–0.742)	0.713 (0.696–0.731)
GBM	0.779 (0.752–0.806)	0.597 (0.561–0.734)	0.849 (0.712–0.898)	0.628 (0.597–0.732)

ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval; RF, random forest; KNN, K-nearest neighbors; LOG, logistics regression; SVM, support-vector machines; XGB, eXtreme Gradient Boosting; GBM, Gradient Boosting Machine.

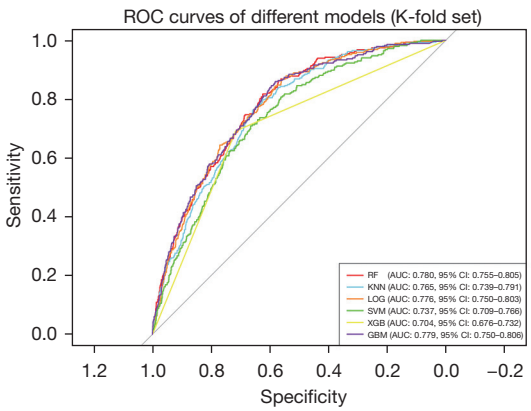

Figure 2 ROC curve for each machine-learning method in the K-fold set. KNN, K-nearest neighbors; SVM, support-vector machine; LOG, logistic regression; RF, random forest; XGB, eXtreme Gradient Boosting; GBM, Gradient Boosting Machine; AUC, area under the curve; CI, confidence interval; ROC, receiver operating characteristic.

in the fields of anesthesia, perioperative care, and pain medicine, including for the prediction of difficult laryngoscopy views (20), hypotension (21), morbidity (22,23), and the risk of weaning from ventilation (24). The model developed and validated in this study was based on the MIMIC-IV database, which consists of comprehensive and high-quality data. There is currently no analysis based on the MIMIC-IV database for predicting hypoxemia after extubation. A recent study developed a CatBoost model to predict extubation failure in ICUs (25). The definition adopted in that study included the need for noninvasive ventilation (NIV), reintubation, or death within 48 h following extubation. However, that definition of extubation failure included patients without oxygenation problems. In addition, the composition ratio of extubation failure cases between the internal dataset and external dataset was significantly different because of the loose definition of extubation failure.

Supervised machine learning is a suitable and useful learning algorithm type for event and risk prediction. Supervised learning is a task-driven procedure, and it uses 1 or more training algorithms for the prediction of prespecified events. For example, Kendale *et al.* (26) conducted supervised machine-learning predictive analytics for the prediction of postinduction hypotension based on electronic health record data. Although current research has hypothesized that artificial intelligence algorithms have so far not surpassed human performance, artificial intelligence has the ability to quickly and accurately screen large amounts of data and to discover correlations and patterns that cannot be detected by human cognition, making it a valuable tool for clinicians. Based on the characteristics of the data, different algorithms have different advantages. The best algorithms in this research were the LightGBM and RF models.

Gradient boosting is an ensemble machine-learning model that combines weak 'learners' into a strong single learner in an iterative fashion (27). LightGBM is a recent modification to the gradient boosting algorithm. It improves the efficiency and scalability of the algorithm without sacrificing its inherited effective performance. LightGBM has the advantages of having high efficiency, support for parallel training, low random access memory usage, high accuracy, large-scale data processing capabilities, and support for categorical features. RF is a classic and powerful supervised algorithm that is highly flexible and integrates multiple unrelated decision trees to construct a forest in a random way for regression or classification (28). The larger

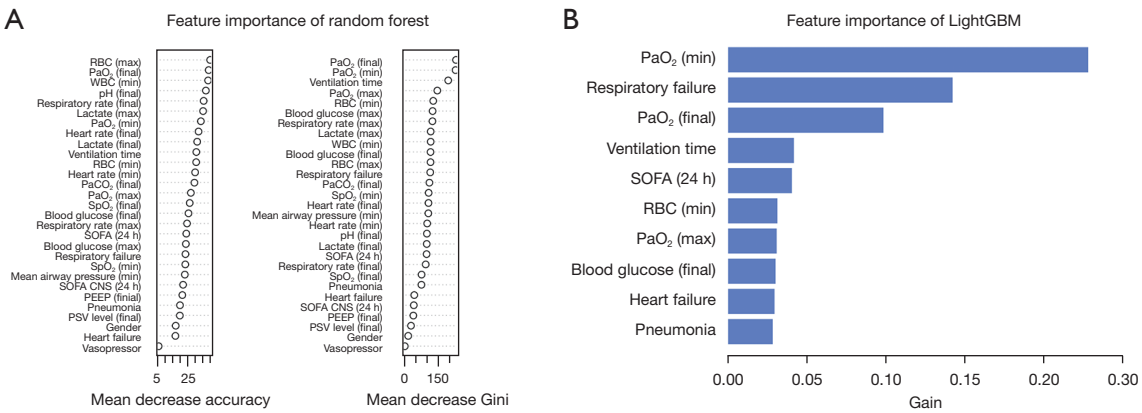

Figure 3 Final feature selection after recursive feature elimination. (A) Feature importance of the random forest model; (B) feature importance of the LightGBM model. LightGBM, Light Gradient Boosting Machine.

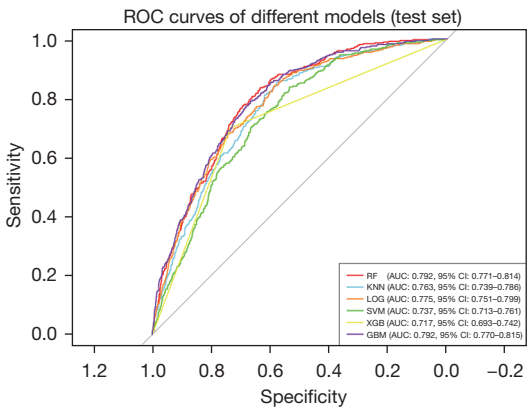

Figure 4 ROC curve for each machine-learning method in the test set. KNN, K-nearest neighbors; SVM, support-vector machine; LOG, logistic regression; RF, random forest; XGB, eXtreme Gradient Boosting; GBM, Gradient Boosting Machine; AUC, area under the curve; CI, confidence interval; ROC, receiver operating characteristic.

the number of decision trees, the stronger the robustness and the higher the accuracy of the RF algorithm. However, this algorithm is more prone to overfitting effects, and its efficiency is lower than that of LightGBM.

Twenty-seven features were included in the feature importance of LightGBM. The most important features included PaO_2 (minimum), respiratory failure, PaO_2 (final), ventilation time, and the SOFA score (24 h). These results were consistent with other studies (29,30). Torrini *et al.* (30) conducted a meta-analysis, and the results indicated that history of respiratory disease, duration of mechanical ventilation, and a lower PaO_2/fraction of inspired oxygen (FiO_2) ratio had the strongest association with extubation outcome. Xie *et al.* (29) conducted a retrospective study, and the results showed that a lower PaO_2/FiO_2 ratio, long duration of mechanical ventilation, and high SOFA score had the strongest association with extubation outcome. Most research results show that a lower PaO_2/FiO_2 ratio before extubation is one of the most important risk factors for hypoxemia after extubation. However, PaO_2 and FiO_2 are 2 independent variables in the MIMIC-IV database, and it is almost impossible to obtain the PaO_2/FiO_2 ratio. A low PaO_2 level indicates poor oxygenation in patients. After weaning from mechanical ventilation and extubation, such patients may experience severe deoxygenation (31). Patients with a long mechanical time tend to have more severe disease. In addition, a long mechanical ventilation time is associated with complications, including ventilator-associated pneumonia and ventilator-induced lung injury (32), which may increase the extubation risks. Other important features included red blood cells (RBCs) (minimum), PaO_2 (maximum), blood glucose (final), heart failure, and pneumonia.

In the sensitivity analyses, all the models with different definition of hypoxemia after extubation, especially those using RF, LOG, and LightGBM, demonstrated acceptable discrimination. These models will further help patients by reducing the incidence of related complications after extubation. For patients, severe hypoxemia is fatal, and it is very helpful for clinicians to accurately predict the occurrence of hypoxemia. The models without multiple imputation, including those using RF, LOG, KNN, and LightGBM, also demonstrated acceptable discrimination. In addition, the results of the sensitivity analyses indicated the

Table 4 ROC, sensitivity, specificity, and accuracy at the best thresholds in the test set

Variables	AUC (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Accuracy (95% CI)	P
RF	0.792 (0.771–0.814)	0.669 (0.584–0.731)	0.814 (0.737–0.888)	0.686 (0.621–0.734)	<0.001
KNN	0.763 (0.739–0.786)	0.601 (0.563–0.639)	0.838 (0.776–0.886)	0.630 (0.599–0.662)	<0.001
LOG	0.775 (0.751–0.799)	0.606 (0.544–0.763)	0.824 (0.665–0.891)	0.635 (0.585–0.754)	<0.001
SVM	0.737 (0.713–0.761)	0.568 (0.521–0.681)	0.803 (0.684–0.870)	0.599 (0.561–0.685)	<0.001
XGB	0.717 (0.693–0.742)	0.736 (0.719–0.752)	0.699 (0.652–0.745)	0.731 (0.715–0.746)	<0.001
GBM	0.792 (0.770–0.815)	0.672 (0.584–0.734)	0.801 (0.718–0.883)	0.687 (0.618–0.736)	<0.001

AUC, area under the curve; CI, confidence interval; RF, random forest; KNN, K-Nearest neighbors; LOG, logistics regression; SVM, support-vector machines; XGB, eXtreme Gradient Boosting; GBM, Gradient Boosting Machine.

Table 5 ROC, sensitivity, specificity, and accuracy at the best thresholds in the sensitivity analyses (different definition of hypoxemia after extubation)

Variables	AUC (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Accuracy (95% CI)
RF	0.780 (0.748–0.812)	0.726 (0.615–0.827)	0.704 (0.582–0.816)	0.724 (0.625–0.813)
KNN	0.760 (0.728–0.793)	0.565 (0.532–0.687)	0.852 (0.714–0.903)	0.584 (0.554–0.691)
LOG	0.778 (0.748–0.808)	0.629 (0.605–0.715)	0.832 (0.730–0.883)	0.642 (0.620–0.717)
SVM	0.729 (0.692–0.765)	0.653 (0.624–0.807)	0.704 (0.531–0.781)	0.658 (0.629–0.792)
XGB	0.707 (0.672–0.741)	0.770 (0.755–0.786)	0.648 (0.582–0.709)	0.762 (0.747–0.778)
GBM	0.777 (0.745–0.808)	0.682 (0.578–0.796)	0.760 (0.628–0.857)	0.687 (0.595–0.785)

ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval; RF, random forest; KNN, K-nearest neighbors; LOG, logistics regression; SVM, support-vector machines; XGB, eXtreme Gradient Boosting; GBM, Gradient Boosting Machine.

Table 6 ROC, sensitivity, specificity, and accuracy at the best thresholds in the sensitivity analyses (dataset without multiple imputation)

Variables	AUC (95% CI)	Specificity (95% CI)	Sensitivity (95% CI)	Accuracy (95% CI)
RF	0.751 (0.716–0.787)	0.698 (0.526–0.777)	0.709 (0.603–0.857)	0.699 (0.560–0.762)
KNN	0.717 (0.679–0.754)	0.682 (0.511–0.726)	0.698 (0.614–0.852)	0.682 (0.545–0.720)
LOG	0.742 (0.707–0.777)	0.755 (0.512–0.797)	0.656 (0.571–0.847)	0.743 (0.550–0.777)
SVM	0.693 (0.655–0.731)	0.744 (0.449–0.784)	0.593 (0.508–0.841)	0.726 (0.490–0.760)
XGB	0.683 (0.647–0.719)	0.752 (0.731–0.774)	0.614 (0.545–0.683)	0.738 (0.717–0.758)
GBM	0.743 (0.709–0.778)	0.663 (0.478–0.738)	0.730 (0.624–0.884)	0.669 (0.520–0.727)

ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval; RF, random forest; KNN, K-nearest neighbors; LOG, logistics regression; SVM, support-vector machines; XGB, eXtreme Gradient Boosting; GBM, Gradient Boosting Machine.

robustness and flexibility of the machine-learning models.

Although the results are promising, there were some limitations in this study. First, despite the comprehensive and high-quality data of the MIMIC-IV database, this study had inherent limitations and potential interference factors due to the data integrity and homogeneity caused by its retrospective nature. Second, although an AUC of 0.792 demonstrates that there is a reasonably better discrimination, there is still great potential for improvement in the model performance before these models are clinically applied. Many clinical features are not available in the database, and some clinical features are only present in a small number of

cases. For example, some studies have shown that there is a correlation between diaphragmatic movement as assessed by ultrasound and extubation failure (33,34), but this feature was not available in the database. With the availability of other features, the predictive power of machine learning will be further improved. Third, this study was a predictive analysis without external validation, which limits the practicality of this precise model in another setting.

The present study showed that the RF and LightGBM model had better predictive power and efficiency than the other models, and we plan to conduct an external cohort for validation in our medical setting.

Conclusions

In conclusion, our machine learning models have considerable potential for predicting hypoxemia after extubation, which help to reduce ICU morbidity and mortality.

Acknowledgments

Funding: This study was supported by the Clinical Research Program of 9th People's Hospital, Shanghai Jiao Tong University School of Medicine (No. JYLJ202013).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://atm.amegroups.com/article/view/10.21037/atm-22-2118/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-22-2118/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

References

1. Summerer V, Arzt M, Fox H, et al. Occurrence of Coronary Collaterals in Acute Myocardial Infarction and Sleep Apnea. J Am Heart Assoc 2021;10:e020340.
2. Fathi M, Massoudi N, Nooraee N, et al. The effects of doxapram on time to tracheal extubation and early recovery in young morbidly obese patients scheduled for bariatric surgery: A randomised controlled trial. Eur J Anaesthesiol 2020;37:457-65.
3. Artime CA, Hagberg CA. Tracheal extubation. Respir Care 2014;59:991-1002; discussion 1002-5.
4. Rassam S, Sandbythomas M, Vaughan RS, et al. Airway management before, during and after extubation: a survey of practice in the United Kingdom and Ireland. Anaesthesia 2005;60:995-1001.
5. Szeles TF, Yoshinaga EM, Alenca W, et al. Hypoxemia after myocardial revascularization: analysis of risk factors. Rev Bras Anestesiol 2008;58:124-36.
6. Yousefshahi F, Samadi E, Paknejad O, et al. Prevalence and Risk Factors of Hypoxemia after Coronary Artery Bypass Grafting: The Time to Change Our Conceptions. J Tehran Heart Cent 2019;14:74-80.
7. Vabalas A, Gowen E, Poliakoff E, et al. Machine learning algorithm validation with a limited sample size. PLoS One 2019;14:e0224365.
8. Hackshaw A. Small studies: strengths and limitations. Eur Respir J 2008;32:1141-3.
9. Wu Y, Huang S, Chang X. Understanding the complexity of sepsis mortality prediction via rule discovery and analysis: a pilot study. BMC Med Inform Decis Mak 2021;21:334.
10. Pattharanitima P, Thongprayoon C, Kaewput W, et al. Machine Learning Prediction Models for Mortality in Intensive Care Unit Patients with Lactic Acidosis. J Clin Med 2021;10:5021.
11. Zhang C, Fu Z, Bai H, et al. Admission white blood cell count predicts post-discharge mortality in patients with acute aortic dissection: data from the MIMIC-III database. BMC Cardiovasc Disord 2021;21:462.
12. Sayed M, Riaño D, Villar J. Predicting Duration of Mechanical Ventilation in Acute Respiratory Distress

Syndrome Using Supervised Machine Learning. J Clin Med 2021;10:3824.
13. Connor CW. Artificial Intelligence and Machine Learning in Anesthesiology. Anesthesiology 2019;131:1346-59.
14. Suarez-Ibarrola R, Hein S, Reis G, et al. Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. World J Urol 2020;38:2329-47.
15. Deo RC. Machine Learning in Medicine. Circulation 2015;132:1920-30.
16. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
17. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130-9.
18. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 4768-77.
19. Azur MJ, Stuart EA, Frangakis C, et al. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20:40-9.
20. Moustafa MA, El-Metainy S, Mahar K, et al. Defining difficult laryngoscopy findings by using multiple parameters: A machine learning approach. Egyptian Journal of Anaesthesia 2017;33:153-8.
21. Hatib F, Jian Z, Buddi S, et al. Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. Anesthesiology 2018;129:663-74.
22. Gao L, Smielewski P, Czosnyka M, et al. Cerebrovascular Signal Complexity Six Hours after Intensive Care Unit Admission Correlates with Outcome after Severe Traumatic Brain Injury. J Neurotrauma 2016;33:2011-8.
23. Gao L, Smielewski P, Czosnyka M, et al. Early Asymmetric Cardio-Cerebral Causality and Outcome after Severe Traumatic Brain Injury. J Neurotrauma 2017;34:2743-52.
24. Gottschalk A, Hyzer MC, Geer RT. A comparison of human and machine-based predictions of successful weaning from mechanical ventilation. Med Decis Making 2000;20:160-9.
25. Zhao QY, Wang H, Luo JC, et al. Development and Validation of a Machine-Learning Model for Prediction of Extubation Failure in Intensive Care Units. Front Med (Lausanne) 2021;8:676343.
26. Kendale S, Kulkarni P, Rosenberg AD, et al. Supervised Machine-learning Predictive Analytics for Prediction of Postinduction Hypotension. Anesthesiology 2018;129:675-88.
27. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. Ann Statist 2001;29:1189-232.
28. Zhao X, Wu Y, Lee DL, et al., iForest: Interpreting Random Forests via Visual Analytics. IEEE Transactions on Visualization and Computer Graphics 2018;25:407-16.
29. Xie J, Cheng G, Zheng Z, et al. To extubate or not to extubate: Risk factors for extubation failure and deterioration with further mechanical ventilation. J Card Surg 2019;34:1004-11.
30. Torrini F, Gendreau S, Morel J, et al. Prediction of extubation outcome in critically ill patients: a systematic review and meta-analysis. Crit Care 2021;25:391.
31. Meade M, Guyatt G, Cook D, et al. Predicting success in weaning from mechanical ventilation. Chest 2001;120:400S-24S.
32. Curley MA, Wypij D, Watson RS, et al. Protocolized sedation vs usual care in pediatric patients mechanically ventilated for acute respiratory failure: a randomized clinical trial. JAMA 2015;313:379-89.
33. Er B, Simsek M, Yildirim M, et al. Association of baseline diaphragm, rectus femoris and vastus intermedius muscle thickness with weaning from mechanical ventilation. Respir Med 2021;185:106503.
34. Dres M, Similowski T, Goligher EC, et al. Dyspnoea and respiratory muscle ultrasound to predict extubation failure. Eur Respir J 2021;58:2100002.

(English Language Editor: A. Muylwyk)

Cite this article as: Xia M, Jin C, Cao S, Pei B, Wang J, Xu T, Jiang H. Development and validation of a machine-learning model for prediction of hypoxemia after extubation in intensive care units. Ann Transl Med 2022;10(10):577. doi: 10.21037/atm-22-2118

Supplementary

Table S1 Clinical and laboratory variables extracted from database

Variable	Unit	Notes
Age	Year	Patient's age
Gender	Male/female	Patient's gender
Weight	Kg	Patient's weight
Height	Cm	Patient's height
Coronary heart disease	Yes/no	Whether the patient has coronary heart disease?
Heart failure	Yes/no	Whether the patient has heart failure?
HBP	Yes/no	Whether the patient has HBP (hypertension)?
Pneumonia	Yes/no	Whether the patient has pneumonia?
Respiratory failure	Yes/no	Whether the patient has respiratory failure?
Cerebrovascular disease	Yes/no	Whether the patient has cerebrovascular disease?
Renal disease	Yes/no	Whether the patient has renal disease?
Liver disease	Yes/no	Whether the patient has liver disease?
Cancer	Yes/no	Whether the patient has cancer?
DM	Yes/no	Whether the patient has DM (diabetes mellitus)?
WBC (max)	k/μL	Maximum value of white blood cell (from ICU admission to extubation)
WBC (min)	k/μL	Minimum value of white blood cell (from ICU admission to extubation)
WBC (final)	k/μL	Final value of white blood cell (from ICU admission to extubation)
RBC (max)	k/μL	Maximum value of red blood cell (from ICU admission to extubation)
RBC (min)	k/μL	Minimum value of red blood cell (from ICU admission to extubation)
RBC (final)	k/μL	Final value of red blood cell (from ICU admission to extubation)
Lactate (max)	mmol/L	Maximum value of lactate (from ICU admission to extubation)
Lactate (min)	mmol/L	Minimum value of lactate (from ICU admission to extubation)
Lactate (final)	mmol/L	Final value of lactate (from ICU admission to extubation)
Platelet count (max)	k/μL	Maximum value of platelet count (from ICU admission to extubation)
Platelet count (min)	k/μL	Minimum value of platelet count (from ICU admission to extubation)
Platelet count (final)	k/μL	Final value of platelet count (from ICU admission to extubation)
Blood glucose (max)	mg/dL	Maximum value of blood glucose (from ICU admission to extubation)
Blood Glucose (min)	mg/dL	Minimum value of blood glucose (from ICU admission to extubation)
Blood glucose (final)	mg/dL	Final value of blood glucose (from ICU admission to extubation)
PaO_2 (max)	mmHg	Maximum value of PaO_2 (from ICU admission to extubation)
PaO_2 (min)	mmHg	Minimum value of PaO_2 (from ICU admission to extubation)
PaO_2 (final)	mmHg	Final value of PaO_2 (from ICU admission to extubation)
$PaCO_2$ (max)	mmHg	Maximum value of $PaCO_2$ (from ICU admission to extubation)
$PaCO_2$ (min)	mmHg	Minimum value of $PaCO_2$ (from ICU admission to extubation)
$PaCO_2$ (final)	mmHg	Final value of $PaCO_2$ (from ICU admission to extubation)
pH (max)	/	Maximum value of pH (from ICU admission to extubation)

Table S1 *(continued)*

https://dx.doi.org/10.21037/atm-22-2118

Table S1 *(continued)*

Variable	Unit	Notes
pH (min)	/	Minimum value of pH (from ICU admission to extubation)
pH (final)	/	Final value of pH (from ICU admission to extubation)
Heart rate (max)	bpm	Maximum value of heart rate (from ICU admission to extubation)
Heart rate (min)	bpm	Minimum value of heart rate (from ICU admission to extubation)
Heart rate (final)	bpm	Final value of heart rate (from ICU admission to extubation)
MBP (max)	mmHg	Maximum value of mean blood pressure (from ICU admission to extubation)
MBP (min)	mmHg	Minimum value of mean blood pressure (from ICU admission to extubation)
MBP (final)	mmHg	Final value of mean blood pressure (from ICU admission to extubation)
Respiratory rate (max)	min^{-1}	Maximum value of respiratory rate (from ICU admission to extubation)
Respiratory rate (min)	min^{-1}	Minimum value of respiratory rate (from ICU admission to extubation)
Respiratory rate (final)	min^{-1}	Final value of respiratory rate (from ICU admission to extubation)
Temperature (max)	℃	Maximum value of temperature (from ICU admission to extubation)
Temperature (min)	℃	Minimum value of temperature (from ICU admission to extubation)
Temperature (final)	℃	Final value of temperature (from ICU admission to extubation)
SpO_2 (max)	%	Maximum value of SpO_2 (from ICU admission to extubation)
SpO_2 (min)	%	Minimum value of SpO_2 (from ICU admission to extubation)
SpO_2 (final)	%	Final value of SpO_2 (from ICU admission to extubation)
Urine output (24 h)	mL	24 h urine output before extubation
PSV level (max)	cmH_2O	Maximum value of PSV level (from ICU admission to extubation)
PSV level (min)	cmH_2O	Minimum value of PSV level (from ICU admission to extubation)
PSV level (final)	cmH_2O	Final value of PSV level (from ICU admission to extubation)
Mean airway pressure (max)	cmH_2O	Maximum value of mean airway pressure (from ICU admission to extubation)
Mean airway pressure (min)	cmH_2O	Minimum value of mean airway pressure (from ICU admission to extubation)
Mean airway pressure (final)	cmH_2O	Final value of mean airway pressure (from ICU admission to extubation)
PEEP (initial)	cmH_2O	Initial value of positive end expiratory pressure (from ICU admission to extubation)
PEEP (final)	cmH_2O	Final value of positive end expiratory pressure (from ICU admission to extubation)
FiO_2 (initial)	%	Initial value of FiO_2 (from ICU admission to extubation)
FiO_2 (final)	%	Final value of FiO_2 (from ICU admission to extubation)
Respiratory rate set (initial)	min^{-1}	Initial value of respiratory rate set (from ICU admission to extubation)
Respiratory rate set (final)	min^{-1}	Final value of respiratory rate set (from ICU admission to extubation)
Minute volume (initial)	mL	Initial value of minute volume (from ICU admission to extubation)
Minute volume (final)	mL	Final value of minute volume (from ICU admission to extubation)
Tidal volume set (initial)	mL	Initial value of tidal volume set (from ICU admission to extubation)
Tidal volume set (final)	mL	Final value of tidal volume set (from ICU admission to extubation)
Plateau pressure (initial)	cmH_2O	Initial value of plateau pressure (from ICU admission to extubation)

Table S1 *(continued)*

https://dx.doi.org/10.21037/atm-22-2118

Table S1 *(continued)*

Variable	Unit	Notes
Plateau pressure (final)	cmH_2O	Final value of plateau pressure (from ICU admission to extubation)
Vasopressor	Yes/no	Whether vasopressor was administered (from ICU admission to extubation)?
Vasopressor (24h)	Yes/no	Whether vasopressor was administered (24 h before extubation)?
SOFA respiration (24 h)	/	24 h SOFA respiration score before extubation
SOFA coagulation (24 h)	/	24 h SOFA coagulation score extubation
SOFA liver (24 h)	/	24 h SOFA liver score extubation
SOFA cardiovascular (24 h)	/	24 h SOFA cardiovascular score before extubation
SOFA CNS (24h)	/	24 h SOFA CNS score before extubation
SOFA Renal (24h)	/	24 h SOFA Renal score before extubation
GCS	/	GCS score before extubation
SIRS	/	SIRS score before extubation

HBP, hypertension; DM, diabetes mellitus; WBC, white blood cell; RBC, red blood cell; ICU, intensive care units; PaO_2, partial pressure of oxygen; $PaCO_2$, partial pressure of carbon dioxide; MBP: mean blood pressure; PSV, pressure support ventilation; PEEP, positive end-expiratory pressure; FiO_2, fraction of inspired oxygen; SOFA, Sequential Organ Failure Assessment; GCS, Glasgow Coma Scale; SIRS, systemic inflammatory response syndrome.

Table S2 Package and tuning parameters in the study

Model	Package	Tuning	
		Step1	Step 2
RF	*randomForest*	small range gird search for *ntree* (350) and *mtry* (7)	large range gird search for *ntree* (500) and *mtry* (7)
LOG	*base*	None	
XGB	*xgboost*	*nround* (75) and *max.depth* (1)	
KNN	*kknn*	*k* (100)	
SVM	*e1071*	*cost* (1) *gamma* (0.19) and *degree* (1) with package function	
GBM	*lightgbm*	*num leaves* (30) and learn *rate* (0.03)	min data in *leaf* (9)

RF, random forest; LOG, logistics regression; XGB, eXtreme Gradient Boosting; KNN, K-nearest neighbors; SVM, support-vector machines; GBM, Light Gradient Boosting Machine.

https://dx.doi.org/10.21037/atm-22-2118