Machine Learning Algorithms Predict Successful Weaning from Mechanical Ventilation Before Intubation

Hyung-Jun Kim (dr.hjkim@snubh.org)
Seoul National University Bundang Hospital  https://orcid.org/0000-0003-1984-8864

Jinchul Kim
Inha University College of Medicine and Hospital

Yun Kwan Kim
Seers Technology Co.Ltd.

Hyeyeon Kim
Crowdworks Co.Ltd.

Hyojung Jung
National Cancer Center

Soonjeong Koh
Yonsei University Wonju College of Medicine

Yujeong Kim
Yonsei University College of Medicine

Dukyong Yoon
Yonsei University College of Medicine  https://orcid.org/0000-0003-1635-8376

Hahn Yi
Asan Institute for Life Sciences

Article

Keywords: Algorithms, Clinical Decision-Making, Intensive Care Units, Noninvasive Ventilation, Organ Dysfunction Scores

Posted Date: November 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2232223/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Prediction of successful weaning from mechanical ventilation in advance to intubation can facilitate discussions regarding end-of-life care before unnecessary intubation. In this context, we aimed to develop a machine-learning-based model that predicts successful weaning from ventilator support based on routine clinical and laboratory data taken before or immediately after intubation. We used the Medical Information Mart for Intensive Care-IV database, including adult patients who underwent mechanical ventilation in intensive care at the Beth Israel Deaconess Medical Center, USA. Clinical and laboratory variables collected before or within 24 hours of intubation were used to develop machine-learning models that predict the probability of successful weaning within 14 days of ventilator support. Of 23,242 patients, 19,025 (81.9%) patients were successfully weaned from mechanical ventilation within 14 days. We selected 46 clinical and laboratory variables to create machine-learning models. The machine-learning-based ensemble voting classifier revealed the area under the receiver operating characteristic curve of 0.863 (95% confidence interval 0.855–0.870), which was significantly better than that of Sequential Organ Failure Assessment (0.588 [0.566–0.609]) and Simplified Acute Physiology Score II (0.749 [0.742–0.756]). The top features included lactate, anion gap, and prothrombin time. The model’s performance achieved a plateau with approximately the top 21 variables. We developed machine learning algorithms that can predict successful weaning from mechanical ventilation in advance to intubation in the intensive care unit. Our models can aid in appropriate management for patients who hesitate to decide on ventilator support or meaningless end-of-life care.

Introduction

Acute respiratory failure can be caused by various conditions, including pulmonary disease, cardiovascular disease, neuromuscular disorder, or required respiratory support after major surgery. Although invasive mechanical ventilation (MV) is a life-sustaining intervention used to assist or replace spontaneous respiration in patients with acute respiratory failure, the procedure is associated with a risk of severe complications such as ventilator-associated pneumonia, pulmonary edema, and acute respiratory distress syndrome.

Inevitably, a proportion of patients will be unable to recover rapidly from ventilatory support, mandating the use of MV for an extended period. The duration of MV is linearly associated with poor outcomes, and the number of days of ventilation support directly correlates with daily incremental costs and unexpected medical conditions like thromboembolic events and posttraumatic stress disorder. Prolonged MV unavoidably accompanies tracheostomy and life-sustaining care, which is not usually desired by patients. Tracheostomy has its advantages in a lower frequency of laryngeal ulcers, less airway resistance, and ease of management. However, prolonged MV and consequent tracheostomy are unlikely to benefit chronically ill patients with an expected dismal prognosis. It should only be performed if it aligns with the patient’s goals and preferences. The possibility of undergoing a tracheostomy can be a reason for hesitancy to intubate elderly or chronically ill patients.
For these reasons, successful early predictions of whether a patient will undergo prolonged MV can support clinical decision-making in many clinical aspects. Several previous models have been suggested to anticipate prolonged MV or tracheostomy\textsuperscript{6,7}, but they were either short-term predictive models or unrealistic models that did not consider patient death. Given this background, we aimed to develop a thorough machine learning model that can predict the possibility of successful weaning from MV within 14 days after intubation, before undergoing intubation.

Results

Baseline Characteristics

Out of 24,379 patients screened, 23,242 were included in the analysis, and 19,025 (81.9\%) successfully weaned from MV within 14 days (Figure 1). For those who successfully weaned from MV, the duration of MV (days, 1.7±2.7 vs. 8.1±10.6, \(P<0.001\)), age (years, 63.7±15.9 vs. 67.8±15.8, \(P<0.001\)), blood urea nitrogen (mg/dL, 22.4±17.2 vs. 33.1±25.4, \(P<0.001\)), creatinine (mg/dL, 1.2±1.2 vs. 1.7±1.5, \(P<0.001\)), anion gap (mEq/L, 13.7±4.0 vs. 17.3±5.8, \(P<0.001\)), Simplified Acute Physiology Score II (SAPS II) (37.5±13.3 vs. 51.6±16.7, \(P<0.001\)), and Sequential Organ Failure Assessment (SOFA) score (2.7±2.5 vs. 3.9±3.5, \(P<0.001\)) were lower, whereas pH (7.4±0.1 vs. 7.3±0.1, \(P<0.001\)) and bicarbonate level (mEq/L, 23.4±4.2 vs. 21.1±5.9, \(P<0.001\)) were higher compared to the other group of patients (Table 1). The proportions of patients with public insurance (56.5\% vs. 48.7\%, \(P<0.001\)), single marital status (60.1\% vs. 51.2\%, \(P<0.001\)), and admission to a medical care unit (36.3\% vs. 19.1\%, \(P<0.001\)) were lower for those who had been weaned from MV within 14 days. Underlying comorbidities were also less common in these patients (Supplementary Table 1).

Development of the Prediction Models

The area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) values resulting from 5-fold cross-validation are shown in Figure 2. The AUROC values (Figure 2A) of the voting classifier (VC), CatBoost classifier (CBC), random forest classifier (RFC), and regularized logistic regression classifier (RLRC) models for the prediction of successful weaning were 0.863 (95\% confidence interval [CI] 0.855–0.870), 0.862 (95\% CI 0.855–0.869), 0.858 (95\% CI 0.852–0.864), and 0.813 (95\% CI 0.805–0.822), respectively. The two conventional scoring models showed AUROC values of 0.749 (95\% CI 0.742–0.756) for SAPS II and 0.588 (95\% CI 0.566–0.609) for the SOFA score. Figure 2B presents the positive predictive value against sensitivity, with AUPRC values of 0.591 (95\% CI 0.572–0.610), 0.592 (95\% CI 0.576–0.608), 0.583 (95\% CI 0.565–0.601), and 0.481 (95\% CI 0.456–0.506), respectively, in each machine learning model. Detailed descriptions of various performance metrics are presented in Table 2. We also generated a confusion matrix for the cross-validation of the VC model (Supplementary Figure 1).
Feature Importance of Each Variable

We chose the VC model as our representative model. The included variables in the model were ranked according to their information gain, and the top three features were lactate concentration, anion gap, and prothrombin time (Figure 3). To better understand the direction of influence each feature has on this model, the SHapley Additive exPlanations (SHAP) algorithm was implemented for this model to explain for each feature the magnitude and direction of its impact on the outcome prediction. The top risk features included anion gap, age, presence of cardiovascular disease, and blood urea nitrogen concentration. Specifically, a higher value or the presence of a variable indicates a higher chance of failure to wean from MV within 14 days (Figure 4).

Change of Model Performance with Variables

We assessed the performance metrics (Cohen's kappa, AUROC, F1-score, and balanced accuracy) of the VC model according to the number of features included (Figure 5). Each metric's performance was calculated as the variables with the highest feature importance were sequentially added. The model reached its plateau performance in all four metrics with approximately 21 variables.

Discussion

Based on the clinical presentation of patients in the ICU, physicians intuitively determine the possibility of successful weaning from MV before endotracheal intubation. However, expressing these intuitions as numerical values is challenging. In this study, we developed and validated well-performing machine learning-based models that predict successful weaning from MV before or immediately after intubation in critically ill patients with ventilator support. The prediction performance of our model was considerably better than that of conventional prognostic scoring models used for patients who need ICU support.

This is the first study to develop a prognostic model that predicts relatively long-term outcomes (14 days) based on variables within a day of intubation. The model characteristics render it clinically pragmatic and facilitate improved discussions about end-of-life care or prolonged MV with a tracheostomy. Previous efforts to predict the prognosis of patients undergoing MV in ICUs have shown several drawbacks. Clark et al. suggested a model consisting of clinical variables (intubation in the ICU, tachycardia, renal dysfunction, acidemia, elevated creatinine, and decreased HCO$_3^-$ concentration) to identify individuals who may need prolonged MV at the time of intubation$^{6,8}$. However, their model was derived from and validated in relatively small patient populations (99 and 225 patients, respectively), and the AUROC value for prolonged MV prediction was about 0.75. Moreover, patients who died within two weeks of starting MV were excluded from the analysis, leading to a selection bias. Several other studies exist but only predicted mortality$^9$ and short-term outcomes$^{10}$, or they did not consider death in model building$^7$. 
Numerous clinical factors have been proposed in predictive models for patients with MV. The I-TRACH model previously extracted tachycardia, renal dysfunction, acidemia, and a decreased HCO$_3^-$ concentration as the main variables for constructing a scoring system$^8$. In another study that reported the prediction of 30-day mortality, the essential features in the models were Acute Physiology and Chronic Health Evaluation II score, Charlson Comorbidity Index, use of norepinephrine, and base excess$^9$. In our study, lactate level and anion gap were the two most important predictors in the final VC model. Per the findings of this study, several prior reports have emphasized the prognostic importance of lactate level and anion gap. The efficacy of early lactate-guided therapy in ICU patients has been reported$^{11}$. The anion gap, a surrogate for levels of unmeasured anions, has been reported in a meta-analysis as an indicator of mortality in critically ill patients$^{12}$.

We proposed the VC ensemble comprising RLRC, RFC, and CBC as our representative model. Historically, logistic and linear regression models have been used for prognosis tasks in clinical decisions concerning ICU-admitted patients$^{13}$. Still, they are not fit for predictor variables with skewed distributions and tend to overfit. To avoid these shortcomings, more complex modeling approaches have been proposed$^{14}$. First, RLRC does not rely on multivariate normality and equal within-group covariance matrices, but predictions require large-scale sample data for stable outcomes$^{15,16}$. Second, RFC works well on data with several input variables and improves its classification accuracy because it keeps bias low and reduces variance. Still, the interpretation is complex, and evaluation is slow$^{17}$. Third, CBC requires lower computational costs but shows better accuracy than other tree-based models and support vector machines$^{18,19}$. The three models (RLRC, RFC, and CBC) showed a tradeoff between precision and recall$^{20}$; therefore, the use of the VC ensemble method improves performance by reducing the variance component of prediction errors made by the contributing models$^{21}$.

Apart from the thorough development of the ensemble model, our study has its strength due to the relatively high number of included patients (n=23,242) provided by the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, which is a well-established open database derived from an ICU in the USA. This is the first study to establish a machine-learning model to predict the weaning probability based on MIMIC-IV data. Moreover, precise search terms like “intubation/invasive ventilation” and “ventilator type/mode” were utilized to establish a more stringent patient selection and outcome definition. Such criteria can provide an example for the selection processes of intubated patients for future studies using the MIMIC-IV database.

Despite our meaningful findings, there are some inherent limitations to our study. First, some imbalance in patient numbers was noted between those with and without successful weaning from MV (81.9% vs. 18.1%, respectively). Therefore, we presented various performance metrics, such as AUPRC values. Second, although most variables were collected before the time of intubation, some variables were collected within 24 hours after intubation. However, we used variables that are not likely to change dramatically according to intubation (e.g., body temperature, blood urea nitrogen, underlying diseases).
In conclusion, we developed and validated a VC ensemble machine learning model that can effectively predict successful weaning from MV within 14 days before or immediately after intubation. Our study indicates that machine learning algorithms may facilitate clinical decision-making, such as identifying patients more likely to benefit from MV before or immediately after endotracheal intubation. This information can relieve the burden and aid doctors in suggesting appropriate management for patients at risk of endotracheal intubation in the ICU, notably for those who hesitate to decide on ventilator support or meaningless end-of-life care due to advanced age or the presence of several comorbidities.

Methods

Data Source

Data on patients requiring MV were obtained from the MIMIC-IV v1.0 database. MIMIC-IV is a well-known large-scale, single-center (Beth Israel Deaconess Medical Center), an open-access database covering 524,740 admissions of 382,278 patients to this center from 2008 to 2019\textsuperscript{22}. The relevant records include demographic data, ICD-9-CM codes, hourly vital signs and input/output, laboratory tests and microbiological culture results, imaging data, treatment procedures, medication administration, and survival data. The database also provides multiple severity-of-illness scores generated from physiologic and laboratory variables on the first day of each ICU admission. MIMIC-IV has several advantages over its previous version MIMIC-III. The composing data are relatively homogenous because MIMIC-IV contains data entirely sourced from the clinical information system iMDSoft MetaVision, the information of “procedure events,” one of the primary sources of clinical information in ICU, is entirely present, and the number of included patients is more significant.

Selection of Participants

For meticulous patient selection, patients with “Intubation” and “Invasive ventilation” codes appearing at least once in the “procedure event” or “chart event” files were selected. Additionally, patients with “Ventilator type” and “Ventilator mode” codes appearing five times or more within 24 hours after the first code were also included (Supplementary Figure 2). The exclusion criteria were as follows: (i) age <18 or >100 years, patients with (ii) pre-existing tracheostomy, and (iii) missing SOFA score and SAPS II.

Data Collection and Outcome Definition

We collected clinical and laboratory variables recorded before and closest to the initiation of MV. For patients who did not have the values before intubation, the nearest value was obtained within 24 hours of intubation. To minimize the impact of intubation on each variable, we selected variables that are less likely to change dramatically after intubation (Supplementary Appendix 1). Missing values were imputed with the median value, and the missing rates are depicted in Supplementary Figure 3. As a comparator,
two severity-of-illness scores, the SOFA and SAPS II scores, were calculated using the codes from Google's BigQuery database (https://github.com/MIT-LCP/mimic-code/tree/main/mimic-iv/concepts/score). The primary outcome was successful weaning within 14 days of intubation, defined as documented MV discontinuation without death.

Model Development

Several machine learning algorithms were used to develop predictive models, such as RLRC\textsuperscript{16}, RFC\textsuperscript{17}, CBC\textsuperscript{19}, and VC ensembles\textsuperscript{21}. The AUROC, AUPRC, Cohen's kappa, and F1-score in the entire dataset were calculated by five-fold cross-validation for each model, and they are presented as averages with 95% CIs. Likewise, the performance parameters for the SOFA and SAPS II scores were calculated, and the AUROCs of each model were compared using the DeLong test\textsuperscript{23}. The VC model's confusion matrix was presented using Cohen's kappa's maximizing threshold value. To determine the optimal hyperparameter setting, the GridSearchCV library (version 0.22) was used to search multiple optimal parameter values to fit estimators automatically. We calibrated the predictions using isotonic regression to verify that the predicted probability reflected the expected probability of successful weaning\textsuperscript{24}. Finally, to better understand how individual variables impact the outcome prediction, a permutation feature importance and SHAP analysis on the best-performing model was conducted\textsuperscript{17,25}. For model development and validation, Python (3.6.9) and its packages such as NumPy (1.19.5)\textsuperscript{26}, pandas (1.1.5)\textsuperscript{27}, scikit-learn (0.23.2)\textsuperscript{28}, matplotlib (3.3.4)\textsuperscript{29}, seaborn (0.11.2)\textsuperscript{30}, rpy2 (3.4.5) (https://github.com/rpy2/rpy2), scipy (1.5.4)\textsuperscript{31}, and shap (0.41.0)\textsuperscript{25}, as well as R (version 3.4.4)\textsuperscript{32} and its package pROC (1.18.0)\textsuperscript{23}, were used\textsuperscript{33}.

Other statistical considerations

To compare the baseline characteristics, categorical variables were presented as total numbers (percentages) and compared using Fisher's exact test. Continuous variables were presented as means ± standard deviations and compared using the Wilcoxon rank-sum test. All statistical analyses in this study were performed using Google BigQuery, Python (version 3.6.9), and R (version 3.4.4), and p-values <0.05 were considered statistically significant.

Data availability

The codes used in this study are made available at https://github.com/yihahn/ml_predict_prolongedMV for non-commercial use.

Declarations

Acknowledgments
This work was Supported by grant no 00-2019-13-2022-0015 from the SNUBH Research Fund.

**Author contributions**

J.K., YK.K., H.Y., and H-J.K. conceived and planned the study and wrote the manuscript. YK.K., H.K., H.J., S.K., Y.K., D.Y., and H.Y., analyzed the data. All authors contributed to the interpretation of the results. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

**Competing interests**

The authors do not have any conflicts of interest to declare.

**References**


Tables

Table 1. Baseline Characteristics of Patients in the Intensive Care Unit According to Successful Weaning from Mechanical Ventilation within 14 Days
<table>
<thead>
<tr>
<th>Variables</th>
<th>Total (n = 23,242)</th>
<th>Successful weaning (N = 19,025)</th>
<th>Prolonged mechanical ventilation or mortality (n = 4,217)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years</td>
<td>64.4 ± 15.9</td>
<td>63.7 ± 15.9</td>
<td>67.8 ± 15.8</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Male sex</td>
<td>14274 (61.4)</td>
<td>11935 (62.7)</td>
<td>2339 (55.5)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Height, cm</td>
<td>169.7 ± 10.8</td>
<td>170.0 ± 10.8</td>
<td>168.7 ± 10.9</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Weight, kg</td>
<td>83.3 ± 23.2</td>
<td>83.8 ± 22.7</td>
<td>81.0 ± 25.4</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Body temperature, °C</td>
<td>36.6 ± 1.3</td>
<td>36.7 ± 1.2</td>
<td>36.5 ± 1.6</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>WBC count, per 10(^9)/L</td>
<td>13.0 ± 7.5</td>
<td>12.7 ± 6.9</td>
<td>14.3 ± 9.8</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>NL ratio</td>
<td>9.1 ± 10.9</td>
<td>8.3 ± 9.3</td>
<td>12.3 ± 15.7</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Hemoglobin, g/dL</td>
<td>10.4 ± 2.1</td>
<td>10.4 ± 2.1</td>
<td>10.6 ± 2.4</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Platelet count, per 10(^9)/L</td>
<td>190.4 ± 100.9</td>
<td>188.9 ± 97.1</td>
<td>197.4 ± 116.2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>BUN, mg/dL</td>
<td>24.4 ± 19.4</td>
<td>22.4 ± 17.2</td>
<td>33.1 ± 25.4</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Creatinine, mg/dL</td>
<td>1.3 ± 1.3</td>
<td>1.2 ± 1.2</td>
<td>1.7 ± 1.5</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Albumin, g/dL</td>
<td>3.2 ± 0.4</td>
<td>3.2 ± 0.4</td>
<td>3.1 ± 0.5</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Total bilirubin, mg/dL</td>
<td>1.3 ± 3.1</td>
<td>1.1 ± 2.4</td>
<td>2.2 ± 5.2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Prothrombin time, INR</td>
<td>1.5 ± 0.7</td>
<td>1.4 ± 0.6</td>
<td>1.7 ± 1.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>pH</td>
<td>7.3 ± 0.1</td>
<td>7.4 ± 0.1</td>
<td>7.3 ± 0.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Sodium, mEq/L</td>
<td>138.9 ± 4.6</td>
<td>138.9 ± 4.3</td>
<td>138.9 ± 5.9</td>
<td>0.900</td>
</tr>
<tr>
<td>Potassium, mEq/L</td>
<td>4.2 ± 0.7</td>
<td>4.2 ± 0.7</td>
<td>4.3 ± 0.8</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Lactate, mmol/L</td>
<td>2.6 ± 1.9</td>
<td>2.4 ± 1.5</td>
<td>3.7 ± 3.1</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Bicarbonate, mEq/L  | 23.0 ± 4.7  | 23.4 ± 4.2  | 21.1 ± 5.9  | < 0.001  
Anion gap, mEq/L  | 14.4 ± 4.6  | 13.7 ± 4.0  | 17.3 ± 5.8  | < 0.001  
SOFA  | 2.9 ± 2.8  | 2.7 ± 2.5  | 3.9 ± 3.5  | < 0.001  
SAPS II  | 40.0 ± 15.0  | 37.5 ± 13.3  | 51.6 ± 16.7  | < 0.001  
Duration of MV, days  | 2.8 ± 5.7  | 1.7 ± 2.7  | 8.1 ± 10.6  | < 0.001  

Numbers are presented as count (percentile) or median ± standard deviation. Abbreviations: MV, mechanical ventilator; WBC, white blood cell; NL, neutrophil/lymphocyte; BUN, blood urea nitrogen; INR, international normalized ratio; SOFA, Sequential Organ Failure Assessment; SAPS II, Simplified Acute Physiology Score II.

Table 2. Performance Metrics of the Developed Machine Learning Models, along with SOFA score and SAPS II

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Cohen's Kappa</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>VC</td>
<td>0.863 (0.855–0.870)*†</td>
<td>0.591 (0.572–0.610)</td>
<td>0.407 (0.379–0.434)</td>
<td>0.551 (0.534–0.569)</td>
</tr>
<tr>
<td>CBC</td>
<td>0.862 (0.855–0.869)*†</td>
<td>0.592 (0.576–0.608)</td>
<td>0.404 (0.373–0.435)</td>
<td>0.550 (0.532–0.568)</td>
</tr>
<tr>
<td>RFC</td>
<td>0.858 (0.852–0.864)*†</td>
<td>0.583 (0.565–0.601)</td>
<td>0.399 (0.386–0.411)</td>
<td>0.545 (0.537–0.552)</td>
</tr>
<tr>
<td>RLRC</td>
<td>0.813 (0.805–0.822)*†</td>
<td>0.481 (0.456–0.506)</td>
<td>0.335 (0.317–0.354)</td>
<td>0.501 (0.491–0.510)</td>
</tr>
<tr>
<td>SAPSII</td>
<td>0.749 (0.742–0.756)*</td>
<td>0.284 (0.264–0.304)</td>
<td>0.280 (0.263–0.297)</td>
<td>0.451 (0.440–0.462)</td>
</tr>
<tr>
<td>SOFA</td>
<td>0.588 (0.566–0.609)*†</td>
<td>0.438 (0.414–0.462)</td>
<td>0.121 (0.096–0.147)</td>
<td>0.330 (0.311–0.349)</td>
</tr>
</tbody>
</table>

Values are presented as means (95% confidence intervals) calculated from 5-fold cross-validation. Hypothesis tests were conducted to determine whether the AUROC values of the models using machine learning algorithms were equal to those of conventional scores. *P-value<0.001 compared to SOFA score. †P-value<0.001 compared to SAPS II. Abbreviations: AUROC, area under the receiver operating characteristics curve; AUPRC, area under the precision-recall curve; VC, voting classifier; CBC,
Catboost classifier; RFC, random forest classifier; RLRC, regularized logistic regression classifier; SOFA, Sequential Organ Failure Assessment; SAPS II, Simplified Acute Physiology Score II.

**Figures**

**Figure 1**

Flowchart of the patient selection process.

Patients with evidence of endotracheal intubation were identified in the MIMIC-IV database. After careful selection, patients were divided into two groups according to whether they had been successfully weaned from MV within 14 days of intubation or not. Abbreviations: SOFA, Sequential Organ Failure Assessment; SAPS II, Simplified Acute Physiology Score II; MV, mechanical ventilation.
Figure 2

Receiver operating characteristic curves (A) and precision-recall curves (B) of the developed machine learning models to predict successful weaning from mechanical ventilation within 14 days of intubation.

The area under ROC curves and area under PR curves and their 95% confidence intervals for all six models are shown in the legends. For comparability, the random classifier is indicated by a black dashed-dotted line. Abbreviations: VC, voting classifier; AUROC, area under the receiver operating characteristics curve; CBC, Catboost classifier; RFC, random forest classifier; RLRC, regularized logistic regression classifier; SOFA, Sequential Organ Failure Assessment; SAPS II, Simplified Acute Physiology Score II; AUPRC, area under the precision-recall curve.
Important explanatory variables for the ensemble voting classifier model.

The means and standard deviations of the top 21 important explanatory variables for the voting classifier model using 5-fold cross-validation and test set permutation are indicated by bars and error bars, respectively. Abbreviations: BUN, blood urea nitrogen; GCS, Glasgow coma scale; WBC, white blood cell; AIDS, acquired immunodeficiency syndrome.
Figure 4

The Beeswarm summary plot in the SHAP package with a randomly selected test set among a 5-fold cross-validation.

This figure indicates which explanatory variables have the most significant influence on the model's predicted value and the tendency of the model's outcome probability as a function of the original value of an explanatory variable simultaneously. Each colored dot represents one variable value for a patient, for which larger values are presented in red and smaller values in blue. The SHAP value, a simplified and computer-friendly Shapley value, is shown on the horizontal axis and is a surrogate of the contribution of a variable value to the model output. Abbreviations: BUN, blood urea nitrogen; ICU, intensive care unit; GCS, Glasgow Coma Scale; WBC, white blood cell; SHAP, SHapley Additive exPlanations.

Figure 5

Performance metrics of the ensemble voting classifier model according to the addition of explanatory variables.
Cohen's kappa ($A$), AUROC ($B$), F1-score ($C$), and balanced accuracy ($D$) of the ensemble voting classifier are shown as the number of explanatory variables in the training sets in descending order of feature importance coefficients. Mean (dot) and standard deviation (error bar) were computed from the datasets by 5-fold cross-validation. The model reached its plateau performance in all four metrics with approximately 21 variables. Abbreviations: AUROC, area under the receiver operating characteristics curve; RFI, relative feature importance.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineDataSupplement.pdf