Developing an Explainable Machine Learning Algorithm to Predict the Mechanical Ventilation Duration of Patients with ARDS in Intensive Care Units

Zichen Wang
Department of Intensive Care Unit, The First Affiliated Hospital of Jinan University
Luming Zhang
Department of Intensive Care Unit, The First Affiliated Hospital of Jinan University
Tao Huang
Department of Clinical Research, The First Affiliated Hospital of Jinan University
Rui Yang
School of Public Health, Xi'an Jiaotong University Health Science Center
Hao Wang
Department of Statistics, Iowa State University
Haiyan Yin
Department of Intensive Care Unit, The First Affiliated Hospital of Jinan University
Jun Lyu (lyujun2020@jnu.edu.cn)
Department of Clinical Research, The First Affiliated Hospital of Jinan University

Research Article

Keywords: ARDS, mechanical ventilation duration, machine learning

Posted Date: May 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1530938/v1

Abstract

Objective: This study aimed to develop an explainable model for predicting mechanical ventilation (MV) duration in patients with acute respiratory distress syndrome (ARDS) using on a machine learning (ML) approach.

Method: Of the 4443 patients with ARDS in the Medical Information Mart for Intensive Care-IV database, 2702 were selected to construct feature set A (age at admission, BMI, Acute Physiology Score III [APS-III], Sequential Organ Failure Assessment [SOFA] score, and other features at MV initiation), and 2228 patients remained for construct feature sets B (age, APS-III, SOFA score, and remaining features after 24 hours MV) and C (A+B). After feature sets were randomly assigned with 70% in a training cohort and 30% in a testing cohort, tenfold cross-validation was conducted on training cohort to determine the best performing model, which was accessed in the related testing cohort and explained using SHapley Additive exPlanations (SHAP).

Result: The tenfold cross-validation results indicated that the Extreme Gradient Boosting model had the best performance on the training set (root-mean-square error [RMSE]=5.78 days [SD=0.52 days]) among six algorithms. The Bland–Altman plot and paired sample t-test results indicated that the predicted and actual values of the optimal model were consistent, with RMSE=6.85 days. The SHAP results indicated that the three most important features for the model were APS-III, age, and BMI, and there was an obvious effect of the interaction between APS-III and age on the SHAP value.

Conclusion: ML models can accurately predict the MV duration of patients with ARDS in intensive care units. The feature set based at MV initiation had better predictive performance than the feature set at 24 hours after MV.

Introduction

Acute respiratory distress syndrome (ARDS) is a diffuse lung disease caused by inflammatory damage to pulmonary capillary endothelial and alveolar epithelial cells during severe infection, shock, trauma, and burns, which can lead to acute hypoxic respiratory insufficiency or failure (1, 2). Globally, 30–47% of patients in intensive care units (ICUs) are diagnosed with ARDS, and the mortality rate ranges from 35–46%(3). In addition to the treatment of primary disease, the primary goal for patients with ARDS is to correct hypoxemia, in which mechanical ventilation (MV) is the most important means of respiratory support(4, 5). Although treatment interventions are beneficial to patients with ARDS, a prolonged MV time related to ARDS will not only prolong the ICU stay and increase the treatment cost, but also increase the risk of pneumonia caused by conditional pathogens, resulting in a poor prognosis(6, 7).

Early prediction of MV duration is also essential for clinical decisions and care strategies, since it affects the timing of tracheostomy(8), initiation of nutrition(9), intensive glycemic control use(10), or transfer to other long-term ventilation units(11). Intensivists therefore tend to predict MV duration for risk stratification and ICU management. However, the current evidence is inadequate for the accuracy of intensivists making early predictions of MV duration(12), indicating the importance of developing accurate and objective tools for predicting MV duration. With the development of computer power, machine learning (ML)—as a subset of artificial intelligence combined with statistical analysis using computer science—is being widely used in critical care, and has impressive performance(13). We therefore aimed to collect the early features of patients with ARDS in ICUs and develop models based on multiple ML algorithms to predict MV duration.

Method

Data source and feature selection

All data were extracted using Structured Query Language from the Medical Information Mart for Intensive Care (MIMIC)-IV database (version 1.0), which contains over 40,000 ICU admissions from 2018 to 2019 at the Beth Israel Deaconess Medical Center. All patient identities were obfuscated to protect privacy following the Health Insurance Portability and

Accountability Act(14)(15). Feature selection was based on previous research and the experience of our clinical experts(16–18). Finally, age, BMI, Acute Physiology Score (APS)-III, Sequential Organ Failure Assessment (SOFA) score, partial pressure of oxygen (PO_2), fraction of inspired oxygen (FiO_2), PO_2/FiO_2 ratio, partial pressure of carbon dioxide (PCO_2), and positive end-expiratory pressure (PEEP) were included. We designed three feature sets to fit the ML model: model A included age, BMI, APS-III, SOFA score, and remaining features at MV initiation; model B included age, APS-III, SOFA score, and remaining features 24 hours after MV; and model C comprised models A and B).

Study population

According to the Berlin definition of ARDS (PaO_2/FiO_2 ratio <300 mmHg)(19), 4443 patients with ARDS were included in this study. After excluding patients with MV duration <24 hours (n=665) and missing features (n=1,076), 2702 and 2228 patients were selected to construct feature set A and feature sets B and C, respectively.

Machine learning models and explanation

Feature sets were randomly assigned, with 70% in the training cohort and 30% in the testing cohort. Six supervised ML algorithms (support vector machine, decision trees, bagging, random forest, gradient boosting decision tree, and Extreme Gradient Boosting [XGBoost]) were fitted to the training cohorts of feature sets A, B, and C. The primary assessment of prediction was the root-mean-square error (RMSE) in the ML regression. After parameter tuning with tenfold cross-validation, which means that the data sets were divided into ten parts (nine were used for training, and one was used for ten runs of testing), the combination of the ML algorithm and the feature set with the best predictive performance (least RMSE) after cross-validation was selected as the final model to be compared with the testing cohort and explained (Figure 1).

Figure 1

Algorithm development improves the complexity of the model, such as in the ensemble or deep learning models, which further complicates the interpretation of the model. In response to this problem, Lundberg and Lee(20) proposed the SHapley Additive exPlanations (SHAP) in 2017 as a unified framework to explain predictions. Based on the logic of game theory, SHAP computes and returns the Shapley value, which represents the model prediction as the contribution of the local accuracy of each covariate to the original model. A variable of higher importance will have fewer missing corresponding Shapley values, and therefore a higher Shapley value.

All statistical analyses were performed using the R Project for Statistical Computing (version 4.0.1) environment.

Results

The clinical characteristics of feature sets A, B, and C as well as the results of univariate analyses are listed in Table 1. Compared with feature set A, patients in sets B and C had significantly longer MV durations. Similarly, patients in sets B and C had significantly higher APS-III scores and PEEP at MV initiation. There were no significant differences in age, BMI, SOFA score, PO_2, PO_2/FiO_2, and PCO_2 at MV initiation between the patients in the two sets. After 24 hours of MV, the PEEP and PO_2/FiO_2 of patients in feature sets B and C significantly increased, while PO_2, FiO_2, and PCO_2 significantly decreased.

Table 1

The tenfold cross-validation results of the six ML models are presented in Table 2. The results demonstrated that the model fitted with feature set A had the best predictive power among all feature sets and ML algorithms. The XGBoost-based feature set A had the best predictive performance among all models (RMSE=5.78 days [SD=0.52 days]).

Table 2

XGBoost was finally selected as the best algorithm, and the performance of the models constructed using the three feature sets was verified using the testing cohort (Table 3). Consistent with the tenfold cross-validation results, compared with feature sets B (RMSE=7.17 days) and C (RMSE=7.15 days), the model built using feature set A had the best prediction accuracy on the testing cohort (RMSE=6.85 days). The Bland–Altman plot and paired-sample t-test results indicated no significant difference between the XGBoost prediction results and the actual results of the testing cohort (Figure 2).

Table 3, Figure 2

The results of SHAP interpretability feature importance for the best performance model was showed in (Figure 3). The results indicated that APS-III score was the most important feature in the model, while the FiO_2 value at MV initiation was the least important. The relationship between each feature and the SHAP value is shown in (Figure 4). Figure 4 also presents the relationship between features, the most interacted features, and the SHAP value. According to Figure 4, the relationships between the three most important features (APS-III score, age, and BMI) and SHAP value were more prominent, while correlation tests indicated that all features were significantly correlated with the SHAP value. When an APS-III score was less than 120, it was positively correlated with the SHAP value; when an APS-III score exceeded 120, the explanatory power of this feature to the model began to decline. Similarly, the peak SHAP value for age was around 33, and the SHAP value then gradually decreased. BMI and SHAP presented a similar W-shaped relationship, reaching peaks at 25, 40, and 50. SHAP also reflected a clear interactive relationship between age and APS-III score. When APS-III scores were lower, older age explained the model to a greater degree. For those younger than 60 years, higher APS-III was associated with a higher SHAP value, and for those older than 60 years, a higher APS-III was associated with a lower SHAP value. There were no obvious interaction effects between other features and the SHAP value.

Figure 3, Figure 4

Discussion

Our results indicate that feature set A consisting of age at admission, BMI, APS-III, and related parameters at MV initiation for patients with ARDS in ICUs was more effective in predicting MV persistence than the feature sets collected 24 hours after MV (feature sets B and C). The optimal model based on XGBoost had a similar prediction accuracy between the training and testing cohorts (RMSE: 6.85 − 5.78 = 1.07 days). Some readily available clinical features collected at MV initiation can accurately predict MV duration, which is very convenient for clinicians in formulating treatment plans. The SHAP results indicated that APS-III scores, which reflects disease severity, occupy the most important position in models for predicting MV duration, and they interacted with age.

Previous studies have suggested that prolonged MV is significantly associated with ICU mortality risk(21–23), ICU readmission risk(24), high ICU hospitalization costs(7, 24), and decreased long-term quality of life(21). Accurate MV duration predictions can therefore allow better risk stratification of patients, assist clinical decision-making, and optimize ICU resource allocation, which is of great significance for improving both cost-effectiveness and patient outcomes. Although there has been considerable research and prediction models on prolonged MV duration, because the definition of prolonged MV was not consistent, the performance evaluation of related prediction models is not applicable to all situations(26). Few previous studies have investigated predictions of specific MV duration, and such predictions based on the clinical experience of intensivists are unsatisfactory(12). New prediction tools must therefore be developed. ML has been applied to predict MV duration(26–29). However, previous studies using MV duration as a dependent variable for ML regression modeling only used data from the MIMIC-III data set, which includes patients admitted to ICUs from 2001 to 2012, which may be outdated data that do not reflect current patient situations.

We believe that our study had particular strengths. First, we used the MIMIC-IV database. In addition to fixing the errors of the MIMIC-III, the MIMIC-IV includes patients from 2008 to 2019, which can better represent the actual current situations of patients with ARDS. Second, we used SHAP to explain the 'black box' of the XGBoost algorithm, making the model easier to understand for clinical staff. Third, our results indicate that features at MV initiation have better predict accuracy than features collected 24 hours after MV, which gives our model a higher potential for early MV duration prediction. Our study also had limitations. Its single-center retrospective design means that more evidence from external validation and prospective studies should be obtained in the future. Due to the limitations of the database, we also could not obtain real-time data for some features, so we could only use the values on the first day of ICU admission, which may have reduced the accuracy of our model. We also applied

Conclusion

ML models can accurately predict MV duration in patients with ARDS in ICUs. The feature set based at MV initiation had better predictive performance than the feature set based at 24 hours after MV.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. Institutional review board approval and informed consent were not required in current study because MIMIC-IV research data is publicly available and approved by Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). The author completed online courses and obtained The PhysioNet Credentialed Health Data License to access the data.

Consent for publication

Not applicable.

Availability of data and material

The data were available on the MIMIC-III website at https://mimic.physionet.org/, https://doi.org/10.13026/C2HM2Q. The data in this article can be reasonably applied to the corresponding author

Competing interests

The author states that there are no conflicts of interest related to this report.

Funding

The study was supported by Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization (2021B1212040007).

Author contributions

ZW created the study protocol, performed the statistical analyses, and wrote the first manuscript draft. LZ conceived the study and critically revised the manuscript. TH maintained the database and performed data collection. RY assisted with data collection and manuscript editing. HW assisted the analysis and explain of statistical methods. JL and HY assisted with manuscript revision and data confirmation. All authors read and approved the final manuscript.

Availability of data and material

The data were available on the MIMIC-III website at https://mimic.physionet.org/, https://doi.org/10.13026/C2HM2Q. The author finished all required training and signed related agreement before authorized data availability from *Physionet*. The data in this article can be reasonably applied to the corresponding author

Acknowledgements

Not applicable .

References

1. Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, et al. Incidence and Outcomes of Acute Lung Injury From the Division of Pulmonary and Criti-cal Care Medicine (G [Internet]. Vol. 16, n engl j med. 2005. Available from: www.nejm.org
2. Nieman GF, Andrews P, Satalin J, Wilcox K, Kollisch-Singule M, Madden M, et al. Acute lung injury: how to stabilize a broken lung. Critical care (London, England) [Internet]. 2018 May 24;22(1):136. Available from: https://pubmed.ncbi.nlm.nih.gov/29793554
3. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. JAMA - Journal of the American Medical Association. 2016 Feb 23;315(8):788–800.
4. Bein T, Grasso S, Moerer O, Quintel M, Guerin C, Deja M, et al. The standard of care of patients with ARDS: ventilatory settings and rescue therapies for refractory hypoxemia. Intensive care medicine [Internet]. 2016/04/04. 2016 May;42(5):699–711. Available from: https://pubmed.ncbi.nlm.nih.gov/27040102
5. Papazian L, Aubron C, Brochard L, Chiche JD, Combes A, Dreyfuss D, et al. Formal guidelines: management of acute respiratory distress syndrome. Vol. 9, Annals of Intensive Care. Springer Verlag; 2019.
6. Ayzac L, Girard R, Baboi L, Beuret P, Rabilloud M, Richard JC, et al. Ventilator-associated pneumonia in ARDS patients: the impact of prone positioning. A secondary analysis of the PROSEVA trial. Intensive Care Medicine. 2016 May 1;42(5):871–8.
7. Bice T, Cox CE, Carson SS. Cost and health care utilization in ARDS--different from other critical illness? Seminars in respiratory and critical care medicine [Internet]. 2013/08/11. 2013 Aug;34(4):529–36. Available from: https://pubmed.ncbi.nlm.nih.gov/23934722
8. Terragni PP, Antonelli M, Fumagalli R, Faggiano C, Berardino M, Pallavicini FB, et al. Early vs Late Tracheotomy for Prevention of Pneumonia in Mechanically Ventilated Adult ICU Patients A Randomized Controlled Trial [Internet]. Available from: https://jamanetwork.com/
9. Kreymann KG, Berger MM, Deutz NEP, Hiesmayr M, Jolliet P, Kazandjiev G, et al. ESPEN Guidelines on Enteral Nutrition: Intensive care. Clinical Nutrition. 2006 Apr;25(2):210–23.
10. van den Berghe G, Wilmer A, Milants I, Wouters PJ, Bouckaert B, Bruyninckx F, et al. Intensive insulin therapy in mixed medical/surgical intensive care units: Benefit versus harm. Diabetes. 2006 Nov;55(11):3151–9.
11. Carpenè N, Vagheggini G, Panait E, Gabbrielli L, Ambrosino N. A proposal of a new model for long-term weaning: Respiratory intensive care unit and weaning center. Respiratory Medicine. 2010 Oct;104(10):1505–11.
12. Figueroa-Casas JB, Connery SM, Montoya R, Dwivedi AK, Lee S. Accuracy of early prediction of duration of mechanical ventilation by intensivists. Annals of the American Thoracic Society. 2014;11(2):182–5.
13. Gutierrez G. Artificial Intelligence in the Intensive Care Unit. Critical care (London, England) [Internet]. 2020 Mar 24;24(1):101. Available from: https://pubmed.ncbi.nlm.nih.gov/32204716
14. Yang J, Li Y, Liu Q, Li L, Feng A, Wang T, et al. Brief introduction of medical database and data mining technology in big data era. Vol. 13, Journal of Evidence-Based Medicine. Blackwell Publishing; 2020. p. 57–69.

15. Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu AD, et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. Vol. 8, Military Medical Research. BioMed Central Ltd; 2021.
16. Seneff MG, Zimmerman JE, Knaus WA, Wagner DP, Draper EA. Predicting the Duration of Mechanical Ventilation: The Importance of Disease and Patient Characteristics. CHEST [Internet]. 1996 Aug 1;110(2):469–79. Available from: https://doi.org/10.1378/chest.110.2.469
17. Figueroa-Casas JB, Dwivedi AK, Connery SM, Quansah R, Ellerbrook L, Galvis J. Predictive models of prolonged mechanical ventilation yield moderate accuracy. Journal of Critical Care [Internet]. 2015;30(3):502–5. Available from: https://www.sciencedirect.com/science/article/pii/S0883944115000532
18. Villar J, Pérez-Méndez L, Kacmarek RM. The Berlin definition met our needs: no. Vol. 42, Intensive Care Medicine. Springer Verlag; 2016. p. 648–50.
19. Force* TADT. Acute Respiratory Distress Syndrome: The Berlin Definition. JAMA [Internet]. 2012 Jun 20;307(23):2526–33. Available from: https://doi.org/10.1001/jama.2012.5669
20. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. 2017 May 22; Available from: http://arxiv.org/abs/1705.07874
21. Chelluri L, Im KA, Belle SH, Schulz R, Rotondi AJ, Donahoe MP, et al. Long-term mortality and quality of life after prolonged mechanical ventilation*. Critical Care Medicine [Internet]. 2004;32(1). Available from: https://journals.lww.com/ccmjournal/Fulltext/2004/01000/Long_term_mortality_and_quality_of_life_after.7.aspx
22. Cox CE, Carson SS, Lindquist JH, Olsen MK, Govert JA, Chelluri L, et al. Differences in one-year health outcomes and resource utilization by definition of prolonged mechanical ventilation: a prospective cohort study. Critical care (London, England) [Internet]. 2007;11(1):R9–R9. Available from: https://pubmed.ncbi.nlm.nih.gov/17244364
23. Pranikoff T, Hirschl RB, Steimle CN, Anderson HL, Bartlett RH. Mortality is directly related to the duration of mechanical ventilation before the initiation of extracorporeal life support for severe respiratory failure. Critical Care Medicine [Internet]. 1997;25(1). Available from: https://journals.lww.com/ccmjournal/Fulltext/1997/01000/Mortality_is_directly_related_to_the_duration_of.8.aspx
24. Zilberberg MD, Luippold RS, Sulsky S, Shorr AF. Prolonged acute mechanical ventilation, hospital resource utilization, and mortality in the United States. Critical Care Medicine [Internet]. 2008;36(3). Available from: https://journals.lww.com/ccmjournal/Fulltext/2008/03000/Prolonged_acute_mechanical_ventilation,_hospital.9.aspx
25. Dasta JF, McLaughlin TP, Mody SH, Piech CT. Daily cost of an intensive care unit day: The contribution of mechanical ventilation*. Critical Care Medicine [Internet]. 2005;33(6). Available from: https://journals.lww.com/ccmjournal/Fulltext/2005/06000/Daily_cost_of_an_intensive_care_unit_day__The.13.aspx
26. Rose L, McGinlay M, Amin R, Burns KE, Connolly B, Hart N, et al. Variation in definition of prolonged mechanical ventilation. Respiratory Care. 2017 Oct 1;62(10):1324–32.

Tables

Table.1

The clinical characteristic of patients in feature sets

	Features set A (N=2,702)	Feature set B/C (N=2,228)	p value
MV Duration	7.1(6.4)	7.9(6.7)	<0.001[α]
Age	62.7(15.7)	62.3(15.6)	0.340[α]
BMI	31.5(10.1)	31.6(9.9)	0.463[α]
APS-III Score	74.5(29.2)	78.4(28.7)	<0.001[α]
SOFA Score	2.7(2.9)	2.8(3.0)	0.221[α]
Features at MV begining			
PEEP	7.3(3.5)	7.5(3.6)	0.048[α]
PO2	111.2(58.5)	112.6(60.6)	0.696[α]
FiO2	72.8(24.8)	74.8(24.5)	0.004[α]
PO2/FiO2	170.0(119.9)	165.9(101.9)	0.139[α]
PCO2	46.7(15.5)	46.7(15.7)	0.994[α]
Features after 24 hours MV			
PEEP	/	8.8(4.2)	<0.001[β]
PO2	/	105.2(41.9)	0.014[β]
FiO2	/	53.0(16.7)	<0.001[β]
PO2/FiO2	/	218.2(149.1)	<0.001[β]
PCO2	/	42.4(11.7)	<0.001[β]

α: the p values were calculated by Mann-Whitney U test; β: the p values were calculated by *Wilcoxon* signed-rank *test*

Table.2

Prediction performance for mechanical ventilation duration among machine learning algorithms

Algorithm	MRMSE±SD		
	Feature Set A	Feature Set B	Feature Set C
Support vector machine	6.00±0.66	6.40±0.71	6.35±0.83
Decision tree	5.94±0.63	6.53±0.68	6.25±0.72
Bagging	5.84±0.54	6.20±0.60	6.20±0.76
Random forest	5.86±0.52	6.17±0.65	6.23±0.75
Gradient Boosting Decision Tree	5.80±0.55	6.00±0.60*	5.98±0.79*
XGboost	5.78±0.52*	6.14±0.68	6.15±0.97

RMSE: Root mean square error; SD: Standard deviation

RMSE and SD were calculated from the result of 10-fold cross-validation

*Best performance model in each feature set

Table.3

Prediction performance for mechanical ventilation duration by XGBoost

XGboost	RMSE
Feature Set A	6.85
Feature Set B	7.17
Feature Set C	7.15

RMSE: Root mean square error

RMSE were calculated for the testing cohort

Figures

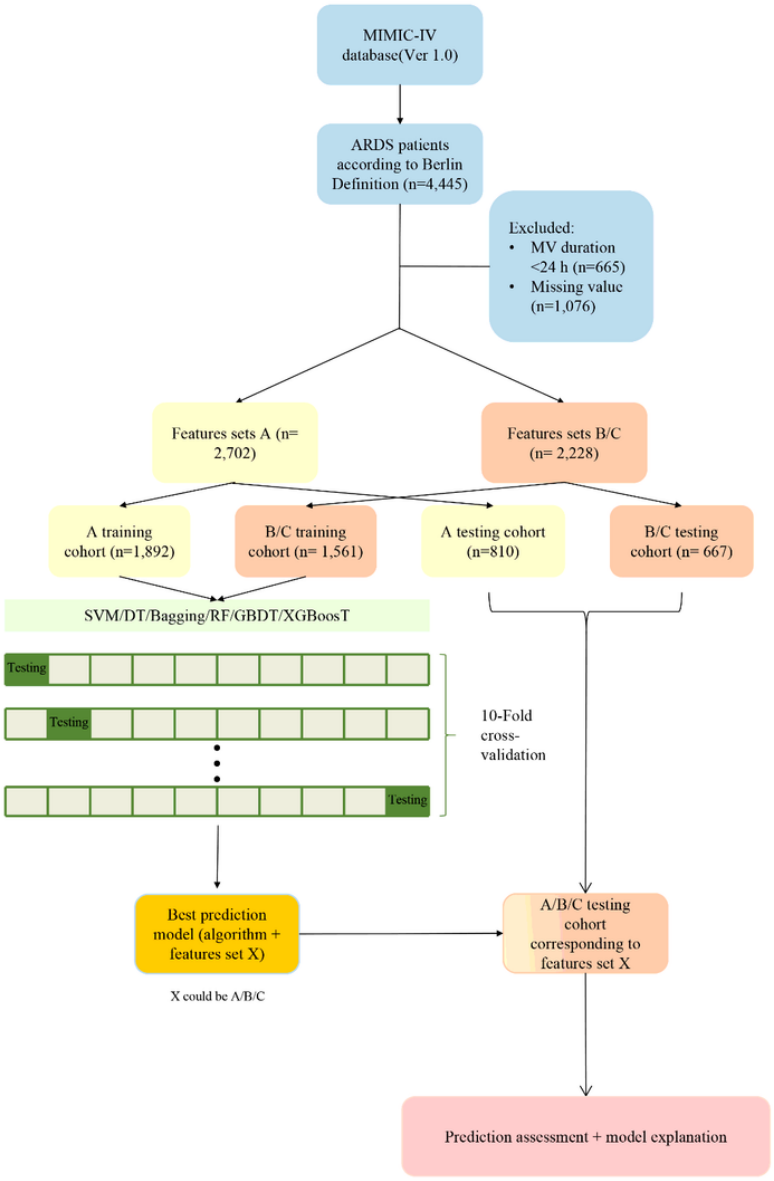

Figure 1

The flow chart of machine learning model construction

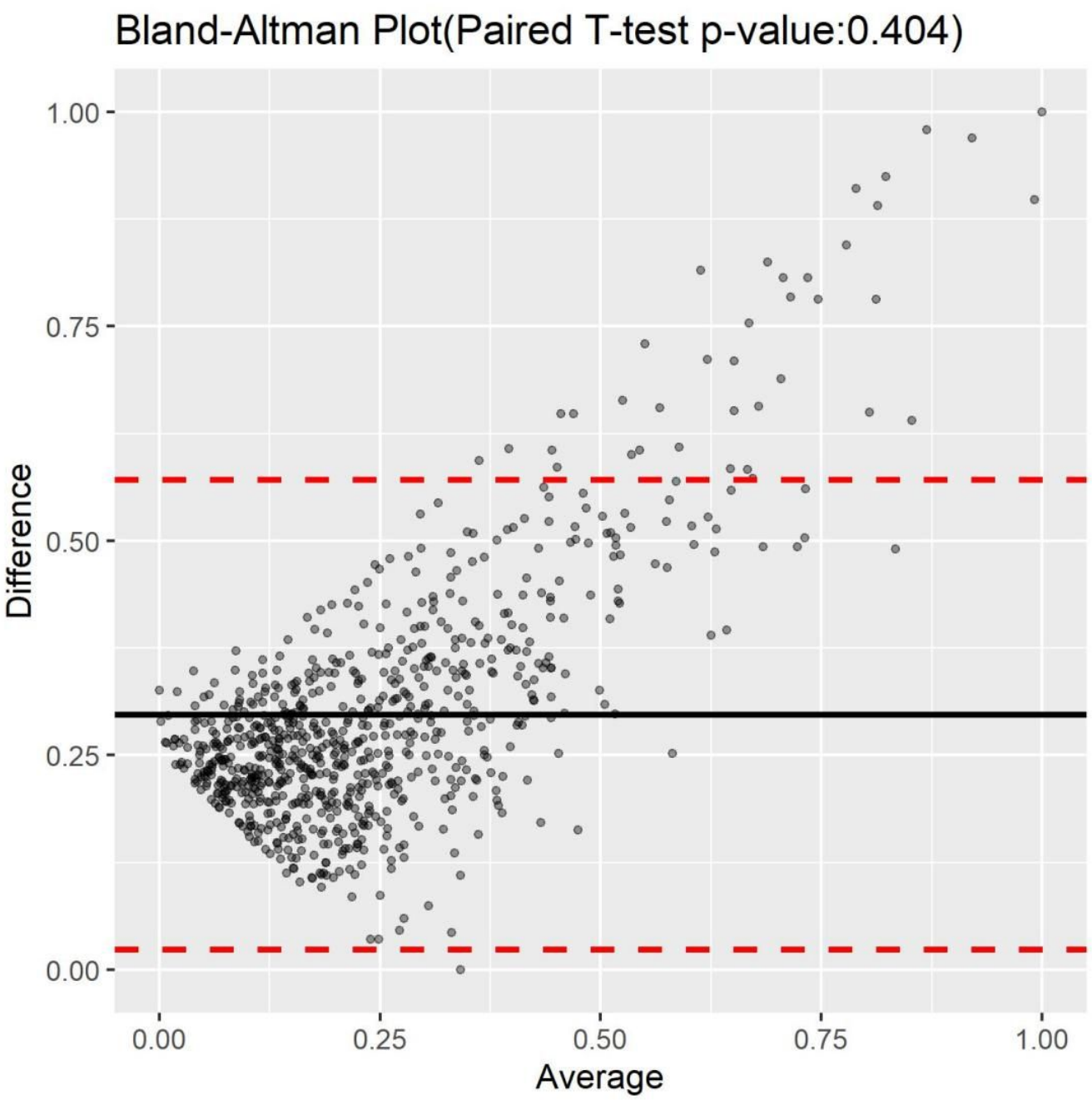

Figure 2

The Bland–Altman plot for actual and prediction value of mechanical ventilation duration by best performing model

The red dotted line represents the 95% limits of agreement

The black implementation represents the average of the differences

The X and Y axes are normalized

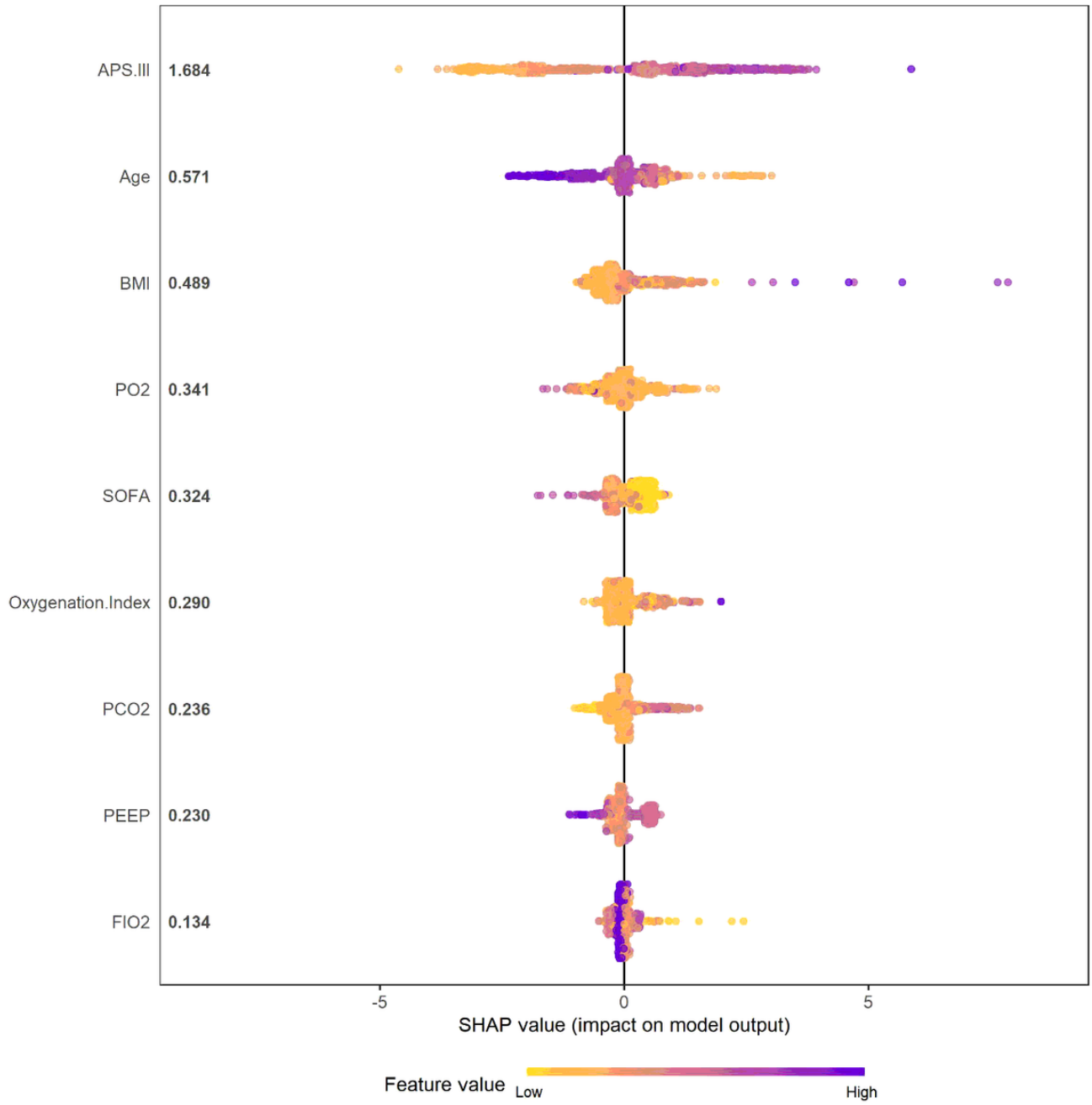

Figure 3

SHapley Additive exPlanations value summary point variable importance plot

Purple points represent high values of the feature, and yellow points represent lower values. A positive SHAP value means that the eigenvalue increases the MV duration.

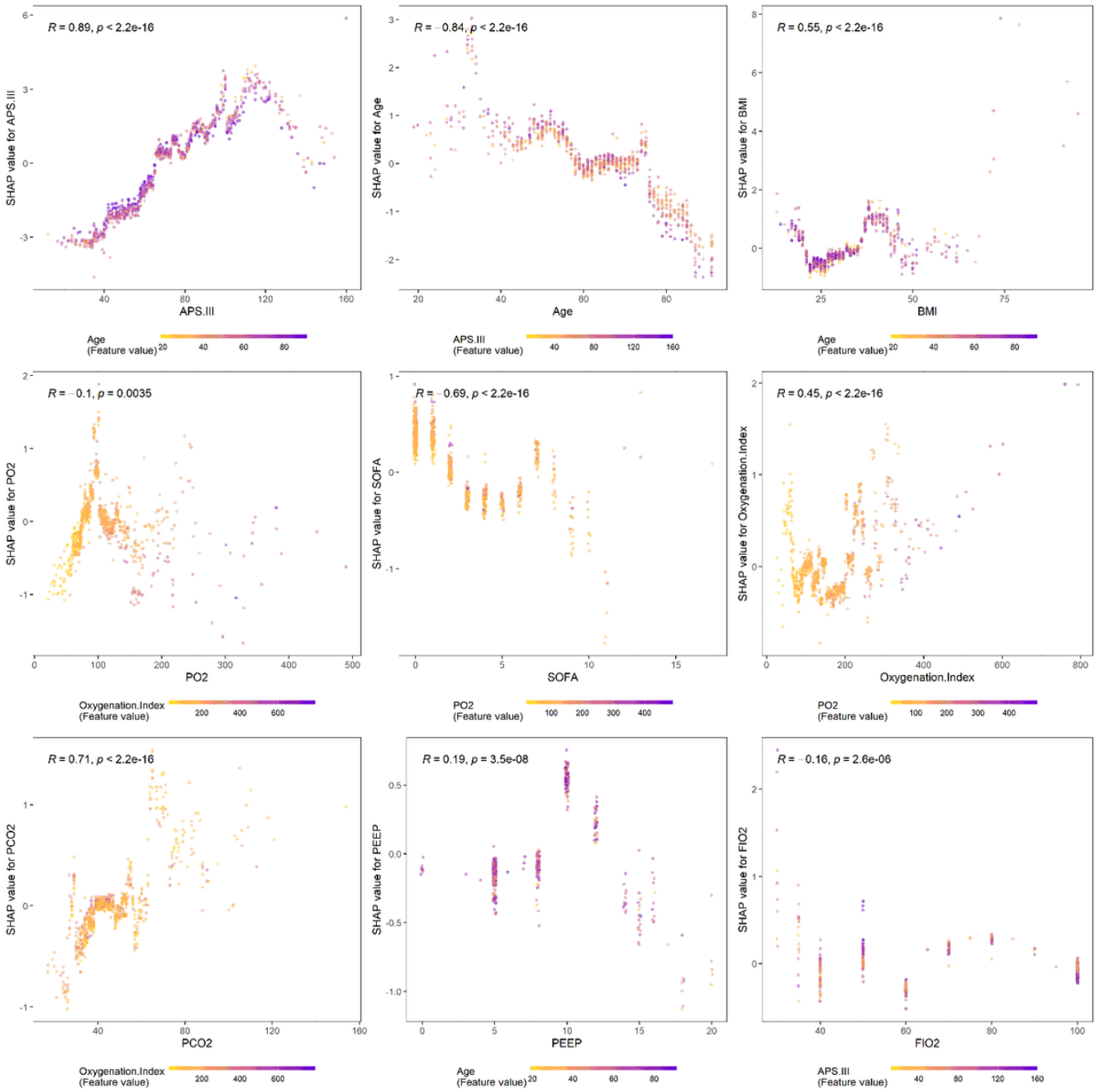

Figure 4

The relationship between feature value and SHapley Additive exPlanations value and feature interactions