Explainable Machine Learning for Predicting Successful COT in Post-Extubation Patients

Chien-Hui Su
Department of Computer Science
National Tsing Hua University
Hsichu, Taiwan
fabienne1023@gapp.nthu.edu.tw

Yen-Hsi Lai
Department of Computer Science
National Tsing Hua University
Hsichu, Taiwan
jessielai0630@gapp.nthu.edu.tw

Cheng-Ya Hsu
Department of Kinesiology
National Tsing Hua University
Hsinchu, Taiwan
xchengya@gmail.com

Jason Yeh
College of Electrical Engineering and Computer Science
National Central University
Taoyuan, Taiwan
jasonjasonyehyeh@gmail.com

Chia-Jung Liu
Department of Internal Medicine, Division of Pulmonology
NTU Hsin-Chu Hospital
Hsichu, Taiwan
m10082100@gmail.com

***Abstract*—Extubation failure remains a significant challenge in critical care, particularly for high-risk patients where resource constraints often limit access to advanced respiratory support such as non-invasive ventilation (NIV) or high-flow nasal cannula (HFNC). This study proposes an explainable machine learning model to predict the success of conventional oxygen therapy (COT) in high-risk patients post-extubation, utilizing data from the MIMIC-IV database. Focusing on a cohort of 10,567 patients meeting traditional high-risk criteria yet managed with COT, we aim to identify key predictors of extubation success—defined as no re-intubation, escalation to HFNC/NIV, or death within 3 days. Multiple machine learning algorithms will be compared for performance and interpretability using techniques like SHAP and PDP to elucidate critical factors. By refining high-risk patient definitions and aligning model insights with clinical standards, this work seeks to enhance decision-making and optimize resource allocation in intensive care settings.**

I. Introduction

Extubation failure is an important issue in critical care, especially among high-risk patients who face increased risks of reintubation, poor outcomes, and a higher mortality rate. Clinical guidelines suggest using non-invasive ventilation (NIV) or high-flow nasal cannula (HFNC) after extubation in these patients. However, in reality, conventional oxygen therapy (COT) is still frequently used, often due to limited medical resources. Although many high-risk patients require advanced respiratory support, some are able to recover successfully with COT alone. This clinical dilemma raises the question: How can we identify high-risk patients who can be safely managed with COT after extubation?

Several studies have shown that machine learning (ML) models can predict extubation outcomes by analyzing clinical data, addressing the limitations of traditional clinical indicators [1]–[3]. This approach may help clinicians make more informed decisions and better allocate limited resources. Previous studies have suggested that HFNC and NIV can help reduce reintubation and mortality rates in high-risk patients after extubation [9]–[11]. However, since many patients still receive COT due to limited resources, it is crucial to develop tools that can identify those who may benefit safely from COT alone.

In addition, the current clinical definition of high-risk patients mainly relies on fixed criteria, such as age or comorbidities, which may not fully reflect a patient's actual condition [8]. In this study, we aim to use ML to build a model that better fits real-world clinical settings, where advanced respiratory support like NIV or HFNC may not always be available. By comparing our model with traditional definitions, we hope to provide a more accurate way to assess which patients truly need advanced support and which may safely recover with COT.

Therefore, we aim to develop an explainable ML-based model to predict the success of COT in high-risk patients, using patient data from the MIMIC-IV database. By identifying the key factors that contribute to the success or failure of extubation, this model could help clinicians make more informed decisions, optimizing patient outcomes and resource allocation.

II. Related Works

A. Machine Learning

Studies showed that ML models could predict extubation outcomes by analyzing clinical data such as vital signs, ventilator parameters, and laboratory results [1], [2]. These models were instrumental in identifying patients at risk for extubation failure, which could guide whether they need NIV or HFNC versus conventional oxygen. The most common models included neural networks, decision trees, and clustering algorithms [5].

Although some previous studies focused on ventilated patients before intubation, they proved that ML models could predict a more accurate success rate than clinical medical indicators. Support Vector Machines (SVM), Neural network (NN), K nearest neighbor (KNN), quadratic discriminant analysis

(QDA), Linear discriminant analysis (LDA), Decision Tree and Logistic Regression, Gaussian Naïve Bayesian, Random Forest, XGBoost, and lightGBM, etc. were compared to predict the successful rate of ventilation methods or intubation rate [6], [7]. To implement ML algorithms in this study, we compared the performance of different ML models to predict the successful rate of conventional oxygen therapy in high-risk patients after extubation.

To select the features in the model training process, previous studies [3] suggested 12 factors significantly associated with extubation failure: age, history of cardiac disease, history of respiratory disease, Simplified Acute Physiologic Score (APACHE) II score, pneumonia, duration of mechanical ventilation, heart rate, Rapid Shallow Breathing Index (RSBI), opposing inspiratory force, lower PaO2 / FiO2 ratio, lower hemoglobin level, and lower Glasgow Coma Scale before extubation . Including these factors in our feature selection process may improve the model's performance and provide insights into analyzing the feature's importance.

B. High-Risk Definition

The traditional clinical medical indicators [8] defined high-risk patients with at least one of the following criteria:

- Age > 65
- APACHE II score > 12 on extubation day
- Body mass index > 30 or a definite diagnosis of obesity
- Inadequate secretions management
- More than one comorbidity, such as heart failure as a primary indication for mechanical ventilation; moderate to severe chronic obstructive pulmonary disease; airway patency problems

Furthermore, previous studies suggested that high-risk patients should use NIV and sometimes HFNC to lower the extubation failure rates or intubation rates, while COT was enough for low-risk patients [9]–[11].

However, in clinical settings, it is essential to allocate resources appropriately to save the greatest number of patients. There is some probability that high-risk patients may be suitable to use COT, while sometimes most hospitals may not have enough resources for NIV and HFNC. Refining the definition of high-risk patients using ML model insights is an urgent issue in clinical.

Researchers have leveraged feature importance analysis and explainable AI techniques, such as Shapley additive explanations (SHAP) plot, partial dependence plot (PDP), and local interpretable model-agnostic explanations (LIME) [1], [3], [4], to identify the most predictive variables, providing a data-driven alternative to redefining subjective clinical thresholds for determining "high-risk" cases. In this study, we employed SHAP and PDP to discern the primary features driving successful COT after extubation prediction.

III. Methods

A. Cohort Selection

In this study, we utilized the Medical Information Mart for Intensive Care (MIMIC-IV) database from the Massachusetts Institute of Technology Laboratory for Computational Physiology as our research target. The cohort selection process is illustrated in Fig. 1.

From 180,733 ICU records, we selected 13,294 patients with the first planned extubation during their first admission to the ICU.

Based on the previous study, patients were excluded if they met any of the following criteria:

- ICU length of stay less than 24 hours ($n = 332$),
- Tracheostomy patients ($n = 126$),
- Mortality within 1 hour of extubation ($n = 9$),
- Give up treatment ($n = 158$).

Patients with a short stay in the ICU were excluded as it indicates that intubation of the patient was required due to a surgical operation.

The inclusion criteria were as follows: patients who received COT instead of HFNC or NIV within 6 hours after extubation.

After applying these criteria, 12,750 patients remained eligible. Among these patients, 2183 of them received HFNC, NIV, or other non-COT respiratory support and were excluded from further analysis. The remaining 10,567 patients who received COT within 6 hours after extubation were grouped as (a) the success group, without observation of deterioration, and (b) the failure group, where patients were reintubated or transferred to other O2 supplement methods (HFNC or HIV) or died within 3 days.

Cohort demographics are detailed in Table I. Notably, a statistically significant difference was observed between the failed and successful extubation groups for both gender (p=0.043) and race (p=0.017). Specifically, males exhibited a higher success rate than failure rate, whereas females showed a higher failure rate than success rate. Similarly, White and Asian patients demonstrated higher success rates compared to their failure rates, while other racial groups presented the opposite trend. These observed demographic disparities might introduce a potential bias in the model's output, a concern that will be further discussed in the Notably, a statistically significant difference was observed between the failed and successful extubation groups for both gender (p=0.043) and race (p=0.017). Specifically, males exhibited a higher success rate than failure rate, whereas females showed a higher failure rate than success rate. Similarly, White and Asian patients demonstrated higher success rates compared to their failure rates, while other racial groups presented the opposite trend. These observed demographic disparities might introduce a potential bias in the model's output, a concern that will be further discussed in section Discussion.

B. Machine Learning Models

We used ML models suggested by previous studies [6], [7], including nine ML (SVM, NN, KNN, Decision tree,

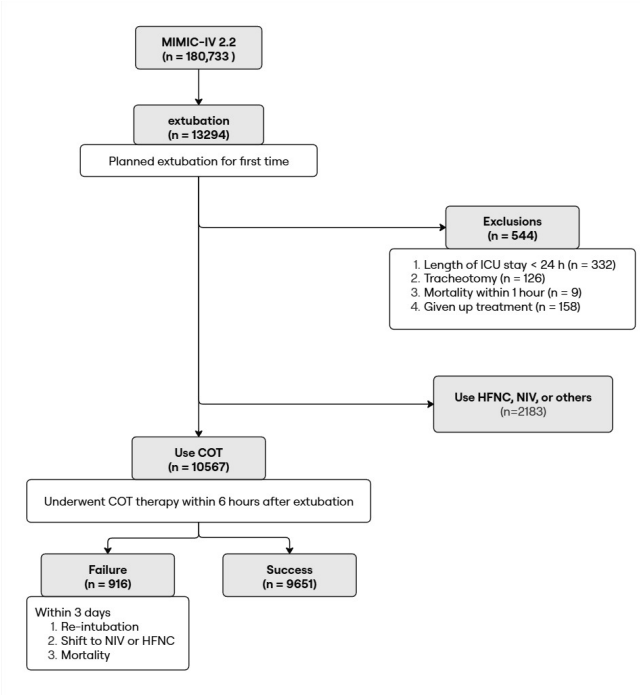


Fig. 1: Cohort selection flow chart.
(HFNC, High-flow Nasal Cannula; NIV, Non-invasive Ventilation; COT, Conventional Oxygen Therapy)

QDA, naive Bayes, LDA, kernel, logistic regression) and six ensemble algorithms (subspace KNN, Bootstrap Random Forest, AdaBooost Tree, GentleBoost Tree, LogitBoost Tree, RUSBooost Tree). Each model underwent hyperparameter optimization and its performance was assessed at its corresponding optimized classification threshold.

C. Feature Selection and Pre-processing

1) relevant features: Following previous studies [3] and expert knowledge, we selected clinically relevant features across several categories before preprocessing: disease severity measures (APACHE III Score), laboratory values (blood gas CO2, O2, hemoglobin level), weaning parameters (RSBI, Pi/e max), physiological indicators (PaO2/FiO2 ratio, Glasgow Coma Scale, heart rate), patient characteristics (age), and clinical history (cardiac disease, respiratory disease, pneumonia, duration of mechanical ventilation). APACHE III score was chosen over APACHE II score for model training since APACHE III score has been shown to be a reliable predictor of hospital mortality in ICU patients [15]. Also, APACHE II represents the traditional standard for high-risk patient identification. To distinguish our methodology from conventional approaches and offer an alternative framework for high-risk classification, we utilized APACHE III, which similarly provides disease severity assessment.

Data for all features were extracted from patient records, prioritizing the closest available measurement within the three days prior to extubation. The APACHE III score, which was calculated by summing the Acute Physiology Score (APS) III score, age points, and chronic health points. The detailed calculation and definition is different from APACHE II and can be refer to [16]. RSBI was calculated as the respiratory rate divided by the spontaneous tidal volume (in liters).

Missing data rates for most of the features ranged from 10% to 20%. However, Pi/e max had a notably higher missing rate at 99%, and RSBI had a missing rate of 80%. Consequently, Pi/e max was excluded from the analysis. RSBI was retained for analysis, as previous studies have demonstrated that multiple imputation via linear regression can effectively handle missing data rates up to 80% [14]. Next, iterative imputation, which implements Multiple Imputation by Chained Equations (MICE) [13], was applied to the remaining 13 features. Outlier detection was conducted using Tukey's method, with outliers defined as values falling more than 1.5 IQR below Q1 or above Q3. These outliers were capped at the respective lower and upper bounds.

Table I presents the statistical analysis of both continuous and categorical variables. All variables, with the exception of pneumonia history, exhibited a p-value below 0.05. The non-significant p-value for pneumonia history was likely attributed to its low prevalence among the patient cohort.

2) High risk features: To compare the performance of ML and traditional methods, we identified high-risk patients from a cohort of 10,576 individuals. The criteria for defining high-risk patients are described in section II-B. Since the MIMIC-IV database does not contain recorded APACHE II scores, we manually calculated the scores using the scoring matrix from the original APACHE II paper [12]. The APACHE II score was the sum of the APS score, age score, and chronic health score.

To calculate the APS score, we used the most recent vital signs and laboratory results recorded within the three days prior to extubation. In total, we considered 24 variables to identify high-risk patients. These include age, BMI, comorbidities (e.g., Obesity, Heart failure, Ischemic heart disease, Pulmonary edema, Arrhythmia, and COPD), core temperature, mean arterial pressure, heart rate, respiratory rate, FiO2, arterial pH, hematocrit, and Glasgow Coma Score, among others. These variables are listed as high-risk features in Table II. As a result, we identified 9,940 high-risk patients among 10,576 subjects.

Table II details the statistical analysis of the high-risk features. Features with p-values greater than 0.05 were excluded from the ablation model training process, including pneumonia of clinically relevant features, due to their lack of a statistically significant difference.

D. Evaluation

The study aims to answer these research Questions:

- Q1: What are the main indicators of success or failure in high-risk patients using COT after extubation?
- Q2: Which ML model demonstrates both high performance and strong explainability?
- Q3: Do the key factors identified by ML models align with clinical definitions of high-risk patients?

TABLE I: Patient Characteristics by Outcome Group After Preprocessing

Variable	Failed (n=916)	Success (n=9651)	P-value
Demographics, n (%)			
Gender			0.043
Male	552 (60.3)	6147 (63.7)	
Female	364 (39.7)	3504 (36.3)	
Race			0.017
White	598 (65.3)	6577 (68.1)	
Unknown	136 (14.8)	1301 (13.5)	
Black	64 (7.0)	612 (6.3)	
Other	53 (5.8)	547 (5.7)	
Hispanic	38 (4.1)	318 (3.3)	
Asian	16 (1.7)	234 (2.4)	
Am. Indian/Alaska	6 (0.7)	15 (0.2)	
Portuguese	5 (0.5)	28 (0.3)	
Continuous variables, mean (SD)			
APACHE III	67.1 (23.2)	55.7 (22.0)	< 0.001
PCO_2	41.0 (7.0)	40.2 (5.9)	< 0.001
PO_2	113.3 (34.6)	125.1 (36.6)	< 0.001
RSBI	65.2 (34.3)	40.2 (29.7)	0.005
P/F ratio	256.0 (85.9)	281.3 (87.2)	< 0.001
Hemoglobin	10.0 (1.5)	10.4 (1.5)	< 0.001
GCS	14.3 (1.7)	14.5 (1.4)	0.002
Age	66.4 (14.9)	65.3 (14.8)	0.037
Vent duration (h)	18.7 (19.0)	16.0 (15.9)	< 0.001
Heart rate	89.5 (17.5)	86.8 (14.7)	< 0.001
Categorical variables, n (%)			
Cardiac disease			0.010
No	503 (54.9)	5731 (59.4)	
Yes	413 (45.1)	3920 (40.6)	
Respiratory disease			< 0.001
No	672 (73.4)	8092 (83.8)	
Yes	244 (26.6)	1559 (16.2)	
Pneumonia			1.000
No	914 (99.8)	9630 (99.8)	
Yes	2 (0.2)	21 (0.2)	

We evaluated and compared ML algorithm performance using AUC-ROC, accuracy, sensitivity (true positive rate), specificity (true negative rate), and Youden index. Sensitivity, specificity, and Youden index were calculated at the optimal threshold. Additionally, model interpretability will be assessed using tools such as PDP and SHAP to visualize the relationship between input features and model outputs, and to identify key contributing factors.

The research will also evaluate the success rate of classifying high-risk patients based on both the clinical definition and the predictions made by ML models, allowing for a comparative analysis of their effectiveness.

An ablation study will be conducted to determine the relative importance of different feature sources, either features suggested by the previous study or from clinical high-risk definitions, that contribute most to accurate predictions.

IV. Results

A. Models Explanation

Among all classifiers, SVM and NN demonstrated superior performance across evaluation metrics (Table III). While SVM achieved higher F1-scores and sensitivity, NN showed better AUC performance and overall balance across metrics. This

TABLE II: High Risk Patient Characteristics by Outcome Group After Preprocessing

Variable	Failed (n=916)	Success (n=9651)	P-value
Continuous variables, mean (SD)			
Temperature	37.1 (0.5)	37.0 (0.5)	< 0.001
Mean artery pressure	82.4 (15.1)	81.3 (14.5)	0.046
Respiratory rate	20.8 (5.6)	19.2 (5.3)	< 0.001
FiO2	0.5 (0.1)	0.5 (0.1)	0.621
A-aDO2	1.6 (0.6)	1.5 (0.6)	< 0.001
PaO2	357.2 (81.9)	368.4 (81.7)	< 0.001
Arterial pH	7.4 (0.1)	7.4 (0.0)	0.001
Serum sodium	139.3 (3.8)	139.1 (3.3)	0.181
Serum potassium	4.2 (0.5)	4.2 (0.5)	0.333
Serum creatinine	1.2 (0.6)	1.1 (0.5)	< 0.001
Hematocrit	30.2 (4.7)	31.0 (4.6)	< 0.001
WBC	12.1 (4.3)	12.2 (4.1)	0.366
APACHE II	9.7 (3.4)	8.7 (3.2)	< 0.001
BMI	28.9 (6.2)	28.5 (5.2)	0.027
Categorical variables, n (%)			
Chronic lung disease			0.007
No	863 (94.2)	9277 (96.1)	
Yes	53 (5.8)	374 (3.9)	
Heart failure			< 0.001
No	790 (86.2)	8832 (91.5)	
Yes	126 (13.8)	819 (8.5)	
Cirrhosis			0.767
No	908 (99.1)	9551 (99.0)	
Yes	8 (0.9)	100 (1.0)	
ERSD			0.596
No	909 (99.2)	9554 (99.0)	
Yes	7 (0.8)	97 (1.0)	
Inadequate secretions management			< 0.001
No	53 (5.8)	1978 (20.5)	
Yes	863 (94.2)	7673 (79.5)	
Obesity			0.002
No	754 (82.3)	8316 (86.2)	
Yes	162 (17.7)	1335 (13.8)	
Ischemic heart disease			0.592
No	567 (61.9)	5881 (60.9)	
Yes	349 (38.1)	3770 (39.1)	
Pulmonary edema			< 0.001
No	860 (93.9)	9434 (97.8)	
Yes	56 (6.1)	217 (2.2)	
Arrhythmia			0.064
No	787 (85.9)	8499 (88.1)	
Yes	129 (14.1)	1152 (11.9)	
COPD			< 0.001
No	793 (86.6)	8694 (90.1)	
Yes	123 (13.4)	957 (9.9)	

TABLE III: Classifier Performance Metrics

Model	AUC	F1	Sensitivity	Specificity	Youden
SVM	0.82	<u>0.96</u>	<u>0.94</u>	0.76	0.70
NN	<u>0.96</u>	0.93	0.87	<u>0.91</u>	<u>0.79</u>
KNN	0.83	0.88	0.81	0.71	0.52
Decision tree	0.79	0.91	0.87	0.61	0.48
QDA	0.83	0.89	0.82	0.72	0.54
Naive Bayes	0.74	0.85	0.77	0.57	0.34
LDA	0.79	0.89	0.83	0.72	0.54
Kernel	0.83	0.92	0.86	0.80	0.66
Logistic	0.80	0.93	0.89	0.75	0.64
Subspace KNN	0.76	0.81	0.70	0.71	0.41
Random Forest	0.89	0.90	0.83	0.81	0.64
Adaboost	0.82	0.88	0.80	0.69	0.50
Gentleboost	0.87	0.85	0.75	0.83	0.59
Logitboost	0.90	0.94	0.90	0.76	0.66
Rusboost	0.82	0.88	0.80	0.69	0.50

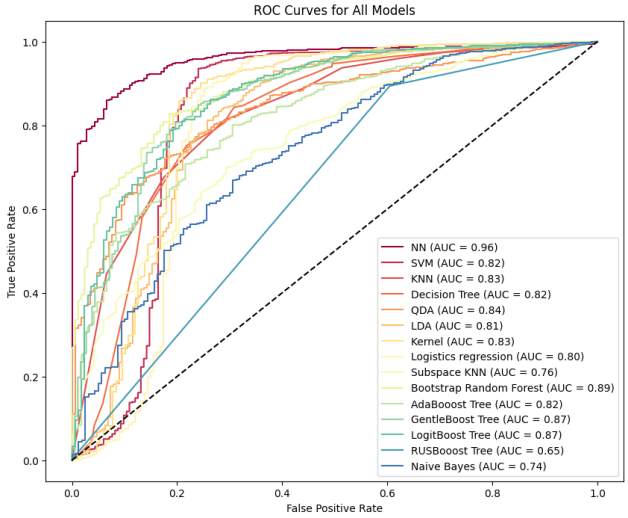


Fig. 2: ROC curves for all models. (NN, Neural Network; SVM, Support Vector Machine; KNN, K Nearest Neighbors; QDA, quadratic discriminant analysis; LDA (linear discriminant analysis))

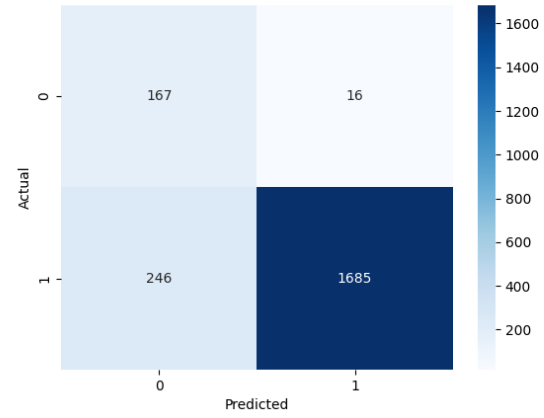
was further supported by the ROC curve (Fig. 2), confirming NN's selection for further analysis.

Figures 3a, 3b, and 3c present the confusion matrix, precision-recall curve, and calibration plot for the NN model, respectively. The confusion matrix confirms balanced prediction performance rather than trivial all-positive classification. The precision-recall curve shows consistently high precision ($\approx$ 1.0) across recall values, demonstrating robust model performance. The original NN calibration curve suggests an underestimation of the success rate. However, post-processing with isotonic regression achieved well-calibrated probability outputs, as shown by the alignment between predicted probabilities and actual positive fractions.

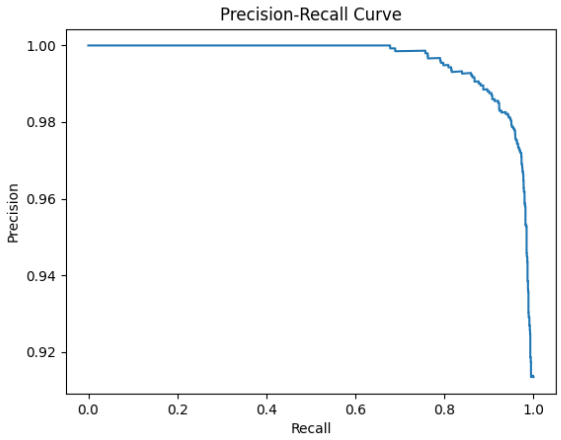
The SHAP analysis conducted by 1000 baground samples (Figure 4a) identified RSBI, P/F ratio, and history of cardiac disease as the top three predictive features. While RSBI followed expected clinical patterns (lower values predict success), P/F ratio and cardiac disease history showed unexpected relationships that contradicted clinical observations, with lower P/F ratios and cardiac disease history associated with success (Fig. 4b). The result indicates that the model may identify underlying feature relationships that remain hidden in traditional statistical approaches.

B. Stratified Analysis

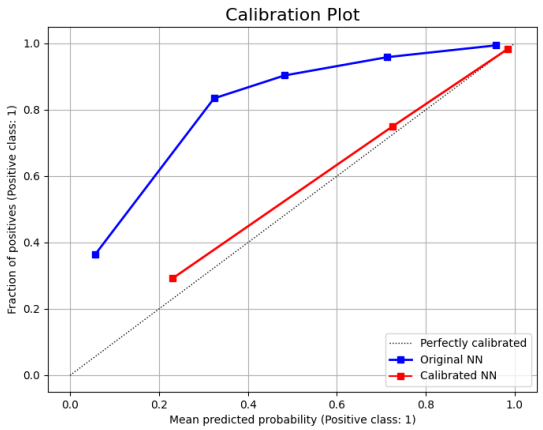
We performed a stratified analysis to investigate the relationship between RSBI, cardiac disease history, and P/F ratio. Table IV indicates that patients with a history of cardiac disease had higher success rates across both high and low RSBI levels. This observation might be explained by the nature of cardiac-related intubations; for instance, some patients intubated due to acute myocardial infarction might recover rapidly after revascularization procedures like cardiac catheter-



(a)

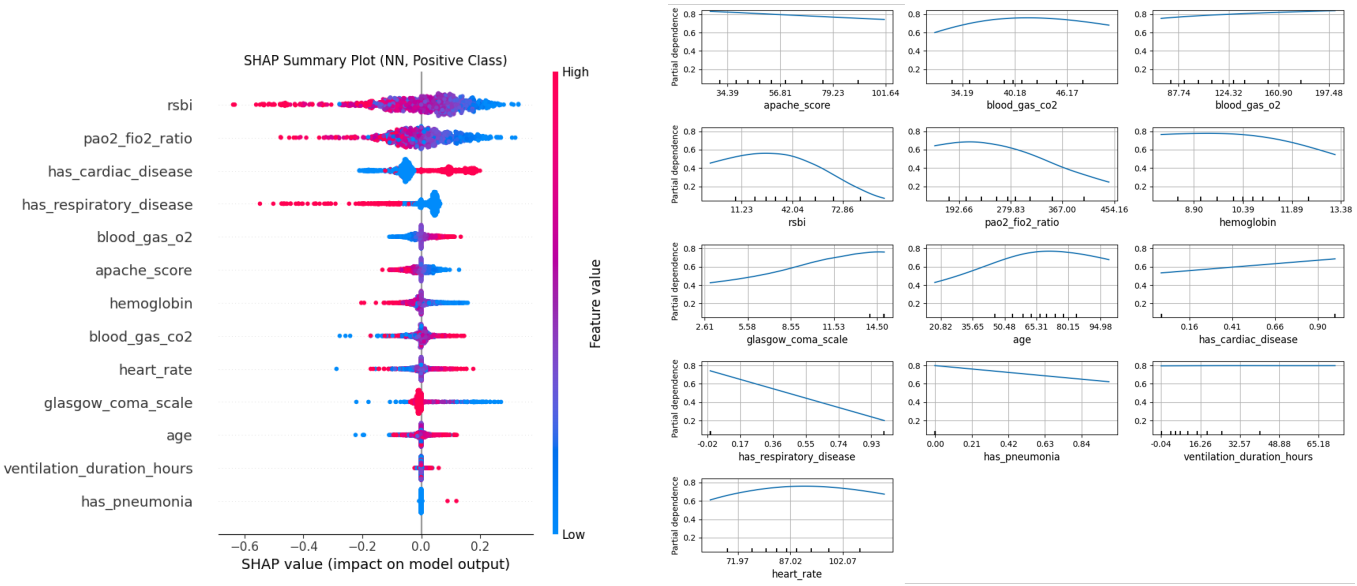


(b)



(c)

Fig. 3: (a) Confusion matrix of NN model predictions. (0: failure, 1: success) (b) Precision-recall curve of NN model predictions. (c) Calibration plot of NN model. (blue: original NN model output, red: NN model output post-processed by isotonic regression, positive class: success)



(a) (b)

Fig. 4: (a) SHAP plot of NN model. Color indicates feature values (blue: low, red: high); horizontal position shows SHAP values (right: positive impact, left: negative impact). (b) PDP plots of NN model. Y-axis shows partial dependence (prediction probability); x-axis shows feature values. Deciles indicate data distribution.

ization, leading to successful COT after extubation. Generally, cardiac conditions, once treated, may show more immediate and predictable improvement in respiratory mechanics and overall physiological stability compared to primary respiratory diseases.

Moreover, Table V indicates that within the high-RSBI group, a lower P/F ratio was associated with successful outcomes. This counterintuitive result suggests a sophisticated interaction among physiological parameters that the NN model effectively discerned. The concurrent findings regarding cardiac disease history and P/F ratio serve to confirm the NN model's proficiency in extracting these implicit relationships between features, thereby providing a robust foundation for subsequent interpretability analyses using methods such as SHAP.

TABLE IV: Success Rate by RSBI Level and Cardiac Disease History

RSBI Level	Cardiac Disease	Success Rate
Low	No	0.958
Low	Yes	0.966
Medium	No	0.951
Medium	Yes	0.883
High	No	0.817
High	Yes	0.873

C. Ablation Study

An ablation study was performed to ascertain the importance of RSBI in the NN model. By excluding RSBI from the feature set during training, we observed a marked reduction in test set performance, as presented in Table VI. Specifically, the AUC fell from 0.96 to 0.71, and the Youden index dropped from 0.79 to 0.27. This strong performance degradation conclusively indicates that the model heavily relies on RSBI as a critical feature, consistent with its established clinical significance.

TABLE V: Count and Mean P/F Ratio Values by RSBI Level and Outcome

RSBI Level	Count		Mean P/F Ratio Values	
	Failure	Success	Failure	Success
Low	136	3386	301.70	328.43
Medium	244	3279	275.96	277.50
High	536	2986	235.40	231.96

We also conducted a comparative analysis between two approaches for training the NN model: one that incorporated the identified "high-risk" features in addition to the clinically relevant features, and another that exclusively used the "high-risk" features. The performance of both these models was found to be inferior to that achieved by using only the clinically relevant features. Notably, the model trained solely on "high-risk" features exhibited the poorest outcome.

TABLE VI: Performance Metrics for Different Feature Selection Methods

Model	AUC	F1	Sensitivity	Specificity	Youden
NN	0.96	0.93	0.87	0.91	0.79
- rsbi	0.71	0.74	0.60	0.67	0.27
+ high-risk	0.70	0.89	0.85	0.39	0.24
only high-risk	0.69	0.88	0.84	0.32	0.16
Traditional definition	-	0.12	0.06	0.99	0.05

D. High-Risk Re-Definition

We present a confusion matrix in Fig. 5 based on patients' high-risk classification and their outcomes after using COT following extubation. As shown, the majority of patients (94.1%) are classified as high-risk. Among the non-high-risk group, 98.7% experienced successful outcomes, which aligns with findings from previous studies. Interestingly, 90.9% of high-risk patients also had successful outcomes. This suggests that many high-risk patients still respond well to COT, indicating that the traditional high-risk criteria may not be sufficient for guiding oxygen therapy decisions.

Table VI compares the performance of various neural network models trained on different feature sets. The model trained on clinically relevant features significantly outperforms both the model trained solely on high-risk indicators and the model trained on a combination of relevant and high-risk features. These findings support the hypothesis that the conventional high-risk definition does not adequately capture patients' true respiratory status. That is, we can redefine high-risk based on the outputs of our explainable model.

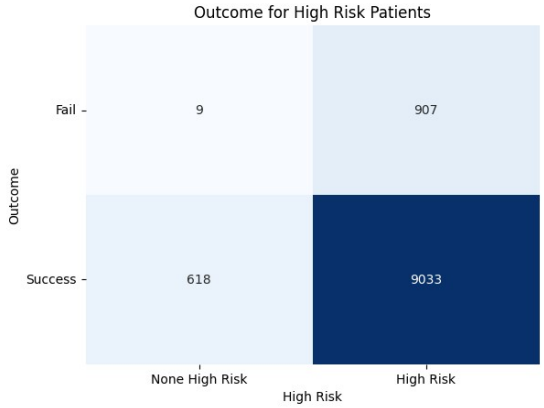


Fig. 5: Outcome for High-risk and Non-high-risk patients

TABLE VII: Model Performance Comparison Across Gender Subgroups

Gender	AUC	F1	Sensitivity	Specificity	Youden
Male	0.97	0.93	0.88	0.92	0.80
Female	0.95	0.92	0.86	0.91	0.76

TABLE VIII: Model Performance Comparison Across Race Subgroups

Race	AUC	F1	Sensitivity	Specificity	Youden
White	0.97	0.93	0.87	0.92	0.79
Black	0.93	0.93	0.87	0.92	0.80
Asian	0.94	0.93	0.88	1	0.88
Others	0.95	0.92	0.87	0.89	0.76

DISCUSSION

This study developed an interpretable ML model to predict the success of COT in high-risk post-extubation patients. While the model achieved strong performance overall, subgroup analysis by sex and race revealed potential sources of bias that may affect its generalizability and fairness.

Gender and Racial Bias

Our analysis revealed that the machine learning model exhibited slight performance differences across gender and racial subgroups. The model performed marginally better in males (AUC 0.97) compared to females (AUC 0.95). This gap may be attributed to sample size imbalance or underlying clinical differences between genders. From a machine learning perspective, the larger proportion of male samples in the training data likely provided the model with more robust examples of male-specific clinical patterns, enhancing prediction accuracy for males. Additionally, clinical features may have different distributions or predictive values between genders, and without sex-specific feature engineering or fairness-aware training, the model may underperform in capturing female-specific variations. Overfitting to the majority male group and absence of bias mitigation strategies can further exacerbate this performance gap.

Regarding race, while the model demonstrated excellent specificity (100%) for the Asian subgroup, this could indicate overfitting due to smaller sample size, raising concerns about generalizability. The model's highest overall AUC was observed in the White subgroup, likely reflecting the largest and most representative sample. Differences in performance among White, Black, and Other groups further highlight the potential for bias, emphasizing the need for more balanced datasets and fairness-aware algorithms to ensure equitable predictive performance across racial groups.

Need for External Validation

Although the model showed excellent performance on the MIMIC-IV dataset, its generalizability to other healthcare systems remains untested. Given variations in patient demographics, clinical practices, and resource availability worldwide, external validation using multi-center and multi-national datasets is critical.

Future work should also incorporate prospective clinical trials and clinician feedback to ensure safe and practical integration into ICU decision-making workflows.

Comparison with Traditional High-Risk Definitions

Traditional criteria (e.g., age > 65, APACHE II > 12, BMI > 30) tend to overclassify patients as high-risk, potentially leading to overuse of advanced respiratory support. Our model identified that approximately 90.9% of traditionally defined high-risk patients still succeeded with COT). By integrating dynamic physiological variables and clinical history, the model more accurately stratifies patient risk, enabling resource optimization.

Clinical Implications and Limitations

The explainable AI approach enhances clinical trust by providing interpretable predictions based on SHAP . It facilitates personalized respiratory therapy decisions, potentially reducing reintubation rates and improving patient outcomes.

Limitations include retrospective data bias, reconstructed APACHE II scores, and lack of real-time dynamic monitoring. Prospective studies and integration of temporal data are warranted.

Conclusion

Clinically, this model addresses the critical challenge of respiratory support allocation in resource-limited ICU settings. By accurately identifying high-risk patients who can safely receive COT without escalation to NIV or HFNC, it helps reduce unnecessary advanced interventions, minimize patient discomfort and complications, and improve ICU throughput and outcomes.

Future research should explore incorporating temporal dynamics and multimodal data, such as imaging and electronic health records, to further enhance predictive accuracy and clinical relevance.

Author Contribution Statements

C.-H.S. (25%), Y.-H.L. (25%), C.-Y. H. (25%), and J.Y (25%) conceived and designed the study. C.-H.S., Y.-H.L., C.-Y. H., and J.Y. implemented the models, did the data analysis, and prepared the materials for the presentation. C.-H.S., Y.-H.L., C.-Y. H. extracted COT extubation cohorts and performed clinical-related feature extraction. J.Y. performed high-risk cohort analysis and feature extraction.

References

[1] Wang, H., Zhao, QY., & Luo, JC. (2022). Early prediction of non-invasive ventilation failure after extubation: development and validation of a machine-learning model. BMC Pulm Med 22, 304. https://doi.org/10.1186/s12890-022-02096-7

[2] Igarashi, Y., Ogawa, K., Nishimura, K., Osawa, S., Ohwada, H., & Yokobori, S. (2022). Machine learning for predicting successful extubation in patients receiving mechanical ventilation. Frontiers in medicine, 9, 961252. https://doi.org/10.3389/fmed.2022.961252

[3] Torrini, F., Gendreau, S., Morel, J., Carteaux, G., Thille, A. W., Antonelli, M., & Mekontso Dessap, A. (2021). Prediction of extubation outcome in critically ill patients: a systematic review and meta-analysis. Critical care (London, England), 25(1), 391. https://doi.org/10.1186/s13054-021-03802-3

[4] Pai, KC., Su, SA., & Chan, MC. (2022). Explainable machine learning approach to predict extubation in critically ill ventilated patients: a retrospective study in central Taiwan. BMC Anesthesiol 22, 351. https://doi.org/10.1186/s12871-022-01888-y

[5] Gallifant, J., Zhang, J., Del Pilar Arias Lopez, M., Zhu, T., Camporota, L., Celi, L. A., & Formenti, F. (2021). Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias. British Journal of Anaesthesia, 128(2), 343–351. https://doi.org/10.1016/j.bja.2021.09.025

[6] Yu, H., Saffaran, S., & Tonelli, R. (2025). Machine learning models compared with current clinical indices to predict the outcome of high flow nasal cannula therapy in acute hypoxemic respiratory failure. Crit Care 29, 101. https://doi.org/10.1186/s13054-025-05336-4

[7] Fu, W., Liu, X., Guan, L., Lin, Z., He, Z., Niu, J., Huang, Q., Liu, Q., & Chen, R. (2024). Prognostic analysis of high-flow nasal cannula therapy and non-invasive ventilation in mild to moderate hypoxemia patients and construction of a machine learning model for 48-h intubation prediction-a retrospective analysis of the MIMIC database. Frontiers in medicine, 11, 1213169. https://doi.org/10.3389/fmed.2024.1213169

[8] Hernández, G., Vaquero, C., Colinas, L., Cuena, R., González, P., Canabal, A., Sanchez, S., Rodriguez, M. L., Villasclaras, A., & Fernández, R. (2016). Effect of Postextubation High-Flow Nasal Cannula vs Noninvasive Ventilation on Reintubation and Postextubation Respiratory Failure in High-Risk Patients: A Randomized Clinical Trial. JAMA, 316(15), 1565–1574. https://doi.org/10.1001/jama.2016.14194

[9] Frat, J., Thille, A. W., Mercat, A., Girault, C., Ragot, S., Perbet, S., Prat, G., Boulain, T., Morawiec, E., Cottereau, A., Devaquet, J., Nseir, S., Razazi, K., Mira, J., Argaud, L., Chakarian, J., Ricard, J., Wittebole, X., Chevalier, S., . . . Robert, R. (2015). High-Flow Oxygen through Nasal Cannula in Acute Hypoxemic Respiratory Failure. New England Journal of Medicine, 372(23), 2185–2196. https://doi.org/10.1056/nejmoa1503326

[10] Rochwerg, B., Einav, S., Chaudhuri, D., Mancebo, J., Mauri, T., Helviz, Y., Goligher, E. C., Jaber, S., Ricard, J. D., Rittayamai, N., Roca, O., Antonelli, M., Maggiore, S. M., Demoule, A., Hodgson, C. L., Mercat, A., Wilcox, M. E., Granton, D., Wang, D., Azoulay, E., … Burns, K. E. A. (2020). The role for high flow nasal cannula as a respiratory support strategy in adults: a clinical practice guideline. Intensive care medicine, 46(12), 2226–2237. https://doi.org/10.1007/s00134-020-06312-y

[11] Oczkowski, S., Ergan, B., Bos, L., Chatwin, M., Ferrer, M., Gregoretti, C., Heunks, L., Frat, J., Longhini, F., Nava, S., Navalesi, P., Uğurlu, A. O., Pisani, L., Renda, T., Thille, A. W., Winck, J. C., Windisch, W., Tonia, T., Boyd, J., . . . Scala, R. (2021). ERS clinical practice guidelines: high-flow nasal cannula in acute respiratory failure. European Respiratory Journal, 59(4), 2101574. https://doi.org/10.1183/13993003.01574-2021

[12] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. (1985) APACHE II: a severity of disease classification system. Crit Care Med, 13(10):818-29. PMID: 3928249.

[13] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40–49. https://doi.org/10.1002/mpr.329

[14] Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M. (2016). Missing Data. In: Secondary Analysis of Electronic Health Records. Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_13

[15] Hsu, C. W., Wann, S. R., Chiang, H. T., Lin, C. H., Kung, M. H., & Lin, S. L. (2001). Comparison of the APACHE II and APACHE III scoring systems in patients with respiratory failure in a medical intensive care unit. Journal of the Formosan Medical Association = Taiwan yi zhi, 100(7), 437–442.

[16] Wagner, D., Draper, E., & Knaus, W. (1989). Chapter 5. Development of APACHE III. Critical Care Medicine, 17(12, Part 2), S199–S203.