Yu et al. Critical Care (2025) 29:101
https://doi.org/10.1186/s13054-025-05336-4

Critical Care

RESEARCH

Open Access

Machine learning models compared with current clinical indices to predict the outcome of high flow nasal cannula therapy in acute hypoxemic respiratory failure

Hang Yu[1], Sina Saffaran[1], Roberto Tonelli[2,3], John G. Laffey[4,5], Antonio M. Esquinas[6], Lucas Martins de Lima[7], Letícia Kawano-Dourado[7], Israel S. Maia[7], Alexandre Biasi Cavalcanti[7], Enrico Clini[2,3*] and Declan G. Bates[1]

Abstract

Background Early identification of patients with acute hypoxemic respiratory failure (AHRF) who are at risk of failing high-flow nasal cannula (HFNC) therapy could facilitate closer monitoring, and timely adjustment/escalation of treatment. We aimed to establish whether machine learning (ML) models could predict HFNC outcome, early in the course of treatment, with greater accuracy than currently used clinical indices.

Methods We developed ML models trained using measurements made within the first 2 h of treatment from 184 AHRF patients (37% HFNC failures) treated at the respiratory ICU of the University Hospital of Modena between 2018 and 2023. For external validation, we used a dataset on 567 AHRF patients (22% failures) comprising 510 patients from the recent RENOVATE trial in Brazil and 57 from the MIMIC-IV and eICU databases in the US. Predictive performance of the ML models was benchmarked against optimized thresholds of the following clinical indices: respiratory rate oxygenation index (ROX) and variants, heart rate to saturation of pulse oxygen (SpO_2) ratio, SpO_2/FiO_2 ratio, PaO_2/FiO_2 ratio, sequential organ failure assessment and heart rate, acidosis, consciousness, oxygenation and respiratory rate scores.

Results Internal and external predictive performance of a Support Vector Machine (SVM) ML model was superior to all clinical indices across all scenarios tested. In external validation on the 567-patient dataset, a SVM model trained on non-invasive measurements had an accuracy of 73%, sensitivity of 73%, specificity of 73%, and AUC of 0.79. The ROX index had an accuracy of 64%, sensitivity of 79%, specificity of 60%, and AUC of 0.74. When arterial blood gasses (ABG's) were also used for model training, the SVM model had an accuracy of 83%, sensitivity of 84%, specificity of 82%, and AUC of 0.82 in external validation on the MIMIC-IV/eICU dataset. The modified ROX index, which requires PaO_2, achieved 70% accuracy, 63% sensitivity, 74% specificity, and AUC of 0.65.

Conclusions Decision support tools based on SVM models could provide clinicians with more accurate early predictions of HFNC outcome than currently available clinical indices. If available, ABG measurements could improve the capability to accurately identify patients at risk of failing HFNC therapy.

*Correspondence:
Enrico Clini
enrico.clini@unimore.it
Full list of author information is available at the end of the article

Introduction

Successful treatment of patients with acute hypoxemic respiratory failure (AHRF) using high-flow nasal cannula (HFNC) therapy has a number of benefits, including improved oxygenation, increased number of ventilator-free days, and reduced risk of intubation and mortality [1–4]. However, recent research also suggests that delayed intubation and mechanical ventilation in patients who fail HFNC could potentially lead to patient self-inflicted lung injury (P-SILI) [5], and increased ICU mortality [6, 7].

Currently, there are no formal guidelines to help clinicians identify patients at risk of HFNC failure [8]. To fill this gap, thresholds based on various clinical indices have been proposed as potential predictors for identifying HFNC failure and determining when to escalate respiratory support. These include heart rate (HR) to saturation of pulse oxygen (SpO_2) ratio [9], respiratory rate oxygenation index (ROX), modified ROX index (mROX) and two variations with heart rate included (ROX-HR and mROX-HR) [10–12], simplified acute physiology and chronic health evaluation II (APACHE II) [13], sequential organ failure assessment (SOFA) score [14], and heart rate, acidosis, consciousness, oxygenation, and respiratory rate (HACOR) [15]. However, significant uncertainty remains regarding the optimal thresholds for these indices, validation of their predictive accuracy on external datasets is rare, and their effectiveness in prompting treatment escalation is limited [16].

A promising approach to address this problem is to employ machine learning (ML) models to find more complex thresholds that can assist in identifying patients who may require closer monitoring or further interventions. However, a recent review of ML models in the context of intensive care medicine noted their low "clinical readiness levels", with only 8 articles (5%) of a total of 172 articles published to date having validated their results on external data, i.e., on data from a source other than the one used for model training [17].

In this study, we developed machine learning models to predict HFNC outcome using data from two previous clinical studies involving 184 AHRF patients carried out at the respiratory intensive care unit of the University Hospital of Modena [18, 19] and performed external validation of the models' predictive performance using independent datasets on 567 AHRF patients (510 from the recent RENOVATE trial in Brazil [20], and 57 from the publicly available MIMIC-IV and eICU databases in the US [21, 22]). We compared the models' performance with that of all currently used clinical indices. In addition, we sought to establish which, and how many, patient measurements were most important in developing models with high predictive accuracy.

Methods

Patient datasets: This retrospective study was performed on the following datasets taken from previously published studies or publicly available databases. The data used for internal training and validation was from two pilot studies carried out at RICU at the University Hospital of Modena, from January 1st, 2018 to May 30th, 2023, and from January 1st, 2021, to June 30th, 2022, respectively [18, 19]. All patients in these studies were adults (age > 18) diagnosed with de novo AHRF and admitted to the RICU after failing conventional oxygen therapy and considered for treatment escalation to HFNC. The first pilot study involved 82 patients (49 successes vs. 33 failures), while the second involved 102 patients (67 successes vs. 35 failures).

For external validation, we used data from the RENOVATE trial [20], a noninferiority, randomized clinical trial which enrolled 883 hospitalized adults with AHRF who received HFNC therapy at 33 hospitals in Brazil between November 2019 and November 2023. To ensure consistency with the training dataset we excluded two patient groups (chronic obstructive pulmonary disease exacerbation with respiratory acidosis, and acute cardiogenic pulmonary edema), but included all patients in the three other groups in the dataset (non-immunocompromised with hypoxemia, immunocompromised with hypoxemia, hypoxemic COVID-19), resulting in 510 patients with 402 HFNC successes vs. 108 failures. The process of study cohort extraction is outlined in Fig. 1a.

As arterial blood gas measurements were not available in the RENOVATE dataset at time points consistent with the training dataset, we extracted additional data from the publicly available MIMIC-IV database from the Beth Israel Deaconess Medical Centre in the United States [21] and the eICU database from 208 hospitals located throughout the United States [22]. We extracted this data using PostgreSQL (*Version 16.0,* Database Management System, https://www.postgresql.org) based on the flow chart illustrated in Fig. 1b to ensure transparency and reproducibility in the cohort selection process—as shown, the key criterion for patient selection was the availability of consistent measurements to those of patients used in training the ML models. This resulted in the selection of an additional 40 patients (21 successes vs. 19 failures) from MIMIC-IV, and 17 patients from eICU (17 successes, 0 failures).

Clinical indices and measurements: Patients' characteristics were collected before or at the start of HFNC (time point T0) and 2 h after HFNC initiation (time point T1). The measurements that were common to all internal and external datasets were: age (y), the SOFA score at T0, the ROX index at T0 and T1, heart rate (bpm) at T0 and T1, respiratory rate (bpm) at T0 and T1, SpO_2

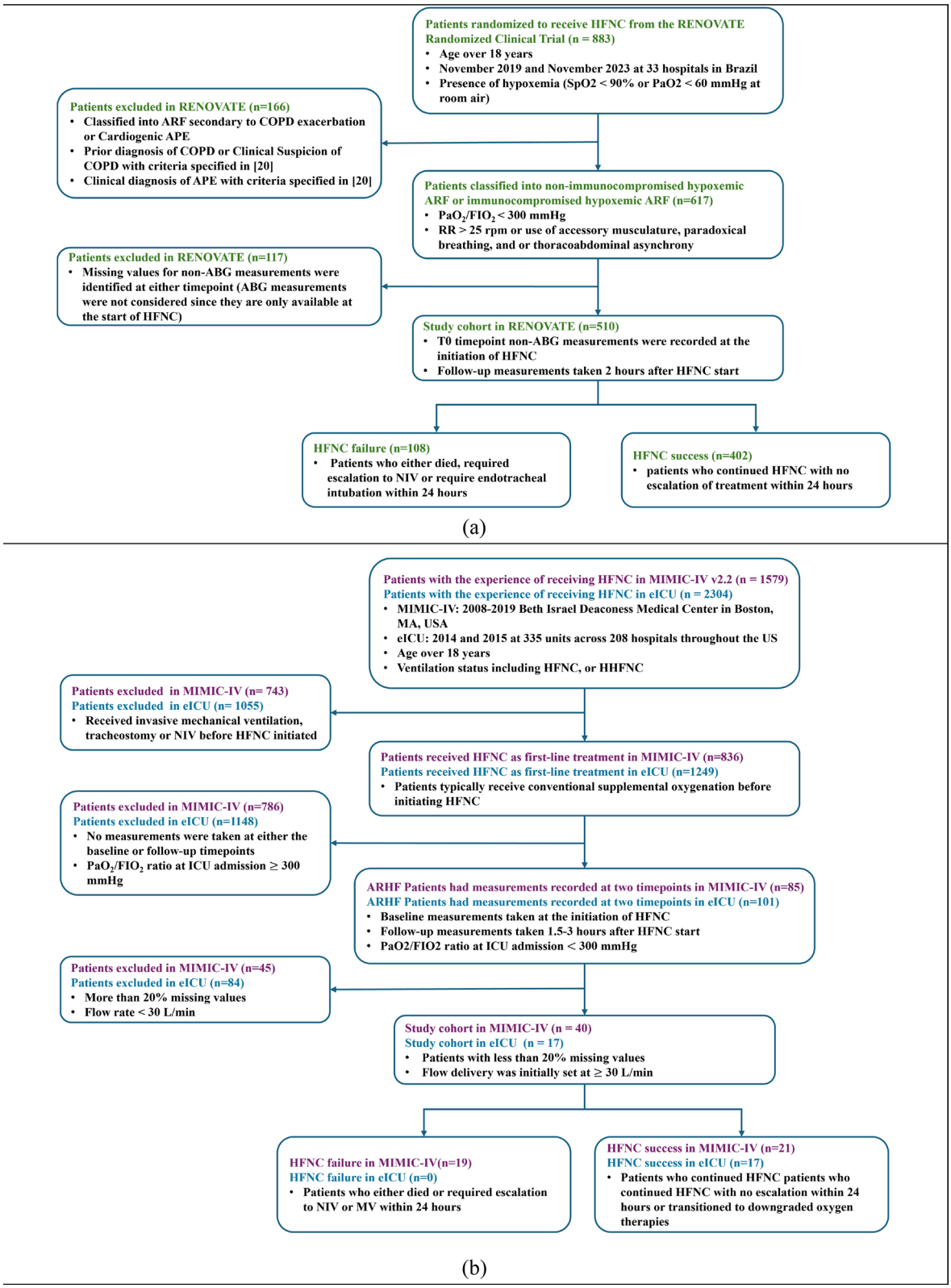

Fig. 1 Flow chart of the data extraction process from RENOVATE and MIMIC-IV/eICU Databases. **a** Study cohort extracted from RENOVATE study **b** Study cohort extracted from eICU/MIMIC-IV database

(%) at T0 and T1, SpO_2/FiO_2 ratio at T0 and T1. PaO_2/FiO_2 ratio (mmHg) at T0 and T1, and partial pressure of carbon dioxide ($PaCO_2$ (mmHg)) at T0 and T1 were also available in the internal and MIMIC-IV/eICU datasets. HFNC failure in all datasets was defined as the need for treatment escalation to non-invasive ventilation (NIV), intubation and mechanical ventilation (MV), or death, within 24 h of HFNC initiation. We tested the predictive accuracy of the following clinical indices: ROX $= \frac{SpO_2}{FiO_2 \cdot RR}$, mROX $= \frac{P_aO_2}{FiO_2 \cdot RR}$, ROX-HR $= \frac{100 \cdot SpO_2}{FiO_2 \cdot RR \cdot HR}$, mROX-HR $= \frac{100 \cdot P_aO_2}{FiO_2 \cdot RR \cdot HR}$, HACOR, SOFA, PaO_2/FiO_2 ratio, SpO_2/FiO_2 ratio and HR/SpO_2 ratio. Optimal thresholds for each index were identified using logistic regression models applied to the training dataset.

Statistical Analysis: Continuous variables are reported as median and inter-quartile ranges (IQR) and compared using non-parametric Mann–Whitney U test and Student's t-test based on the characteristics of the data. The Student's t-test was used for normally distributed data with equal variances, as determined by the Shapiro–Wilk test for normality and Levene's test for equal variances or the Mann–Whitney U test was applied when applicable. Categorical variables were described by counts and frequencies and compared using Fisher's exact test. To calculate the combined P value from the P values of HFNC success and failure between the two cohorts, we applied Fisher's method. All tests were two-sided, and a P value < 0.001 was considered statistically significant.

Development of the machine learning models: We developed and evaluated ML models using the following methods: a range of regression methods including Support Vector Machines (SVM), Decision Tree and Logistic Regression, probabilistic models such as Gaussian Naïve Bayesian, and ensemble methods such as Random Forest, XGBoost, and lightGBM. The outcome-related feature process was performed using a Genetic Algorithm [23]. Hyperparameters were finetuned using Ray Tune with the Asynchronous Successive Halving Algorithm (ASHA) based on distributed GPUs [24]. The optimization score was set to balanced accuracy to prevent biased predictions.

The models were developed and internally validated using two nested cross-validation methods: nested leave-one-out cross-validation (LOOCV) and repeated nested five-fold cross-validation [25–27]. These methods employ a nested cross-validation approach, comprising an inner loop for model selection and an outer loop for model assessment. This design ensures unbiased performance evaluation by maintaining a strict separation between the processes of model selection and assessment, thereby preventing information leakage and reducing the risk of overfitting (see Additional file: Fig. S2). At the external validation stage, models trained on the internal datasets were applied to the data from the previously unseen external datasets to evaluate their generalizability. For sample size estimation using ML models, we employed the simulation-based approach proposed by van der Ploeg T et al., [28]—the minimum sample size required for each model is shown in the (Additional file: Fig. S1). The process for model development, validation, performance analysis, and clinical deployment is outlined in Fig. 2. Software packages used to perform the computations are also detailed in the Additional file.

Evaluation Metrics and Explainability for Predictive Performance: For unbiased evaluation, we considered the following 6 metrics: accuracy, true positive rate (sensitivity), true negative rate (specificity), and area under receiver operating characteristic curve (AUC score). Model calibration was examined in a calibration plot, and discrimination was visually presented using ROC curves. In addition, the decision curve analysis (DCA) curve [29] was applied to assess the clinical utility of the model based on different threshold probabilities for intervention decisions.

The relative importance of different patient measurements in determining predictive performance was computed using shapely additive explanation (SHAP) values [30] and permutation importance (PI) scores [31]. To ensure robust and low variance in the SHAP values, we calculate SHAP values for each patient's features 200 times using repeated nested five-fold cross-validation. The final SHAP values are obtained by averaging all the SHAP values calculated from the validation sets.

Results

Patient characteristics of the internal training dataset: Among the 184 patients in the internal cohort, 68 patients (37%) were identified as HFNC failures and required treatment escalation. As shown in Table 1, at any time, patients who failed HFNC consistently showed higher HACOR score, higher RR, and higher FiO_2 compared with the HFNC success group. In addition, these patients showed lower ROX index scores, SpO_2/FiO_2 and PaO_2/FiO_2. At timepoint T1, group differences were observed in $PaCO_2$, whereas this difference was less evident at the initiation of HFNC. The SOFA score, assessed at or before the start of HFNC, also showed some differences between the two groups. However, no other demographic or clinical characteristics demonstrated significant differences between the two groups.

Predictive performance metrics excluding arterial blood gas (ABG) measurements: Among the developed ML models, the SVM demonstrated the highest balanced accuracy in both the internal and external validation (Additional file: Table S1, Fig. S4). Since ABG

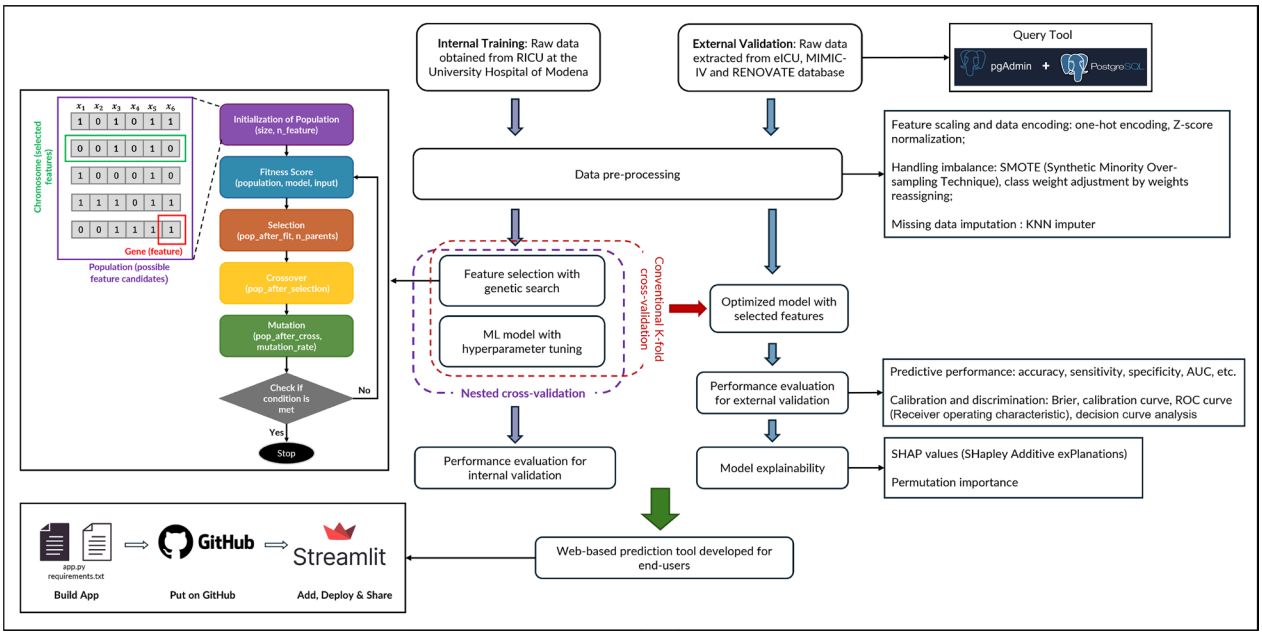

Fig. 2 Process for model development, validation, performance analysis, and deployment

measurements are not typically available in patients receiving HFNC, and are not required for the ROX and ROX-HR indices, we initially trained the SVM model excluding PaO_2, $PaCO_2$ and pH from the dataset. As shown in Table 2, during internal validation with measurements from two time points, the SVM model achieved 90% accuracy, 88% sensitivity, 91% specificity, and 0.93 AUC in LOOCV. Additionally, in nested five-fold cross-validation with 95% confidence intervals, the model had an accuracy of 87% (95% CI: 76-95%), sensitivity of 88% (95% CI: 71-100%), specificity of 89% (95% CI: 76-97%), and an AUC of 0.87 (95% CI: 0.78–0.97). In external validation on the combined Brazilian and US datasets, the model achieved an accuracy of 73%, a sensitivity of 73%, a specificity of 73%, and an AUC of 0.79 (see Fig. 3a).

A Decision Curve analysis (DCA) was conducted to evaluate the utility of the model without using ABG measurements (Fig. 3c). The results indicate that, for thresholds between 15 and 45%, treatment escalation decisions informed by the model provide greater net benefit (the balance between the true positives and the false positives of a particular decision rule or model based on a specified threshold probability) compared to simpler strategies, such as escalating treatment for all patients or not escalating treatment at all. Finally, the calibration curve (Additional file: Figs. S3, S4, and S5) illustrates that the model is well calibrated as evidenced by a clear trend of increasing HFNC failure rates corresponding with higher predicted risk across stratified patient groups. Full results for all the other ML models developed in the study are included in the supplementary material (Additional file: Table S3).

Among the currently available clinical indices, optimized thresholds for ROX and ROX-HR at time T1 found using linear regression models applied to the training dataset performed better than all the others, with ROX achieving 64% accuracy, 82% sensitivity, 60% specificity, and 0.74 AUC in external validation. ROX-HR achieved similar scores of 64% accuracy, 82% sensitivity, 60% specificity, and an AUC of 0.75. The Brier score for ROX and its variants is higher than that of the SVM model, indicating poorer calibration, particularly within intermediate- and high-risk patient groups, where the predicted probability of HFNC failure lies within the intermediate to high probability range (Additional file: Fig. S3). From the DCA (Fig. 3c), the net benefit for ROX(T1) and ROX-HR(T1) under different decision thresholds is similar, only informing meaningful decisions over full intervention or no intervention strategies for thresholds between 10 and 33%. In comparison, the SVM model consistently has a higher net benefit across all thresholds. Although some predictors and indices showed good separation between the success and failure groups in the training set (Additional file: Fig. S9), their distributions exhibited considerable overlap between the two groups, and the optimal cut-off values varied across different datasets (Additional file: Fig. S10). This variability resulted in inconsistent performance across all validation methods for individual indices or predictors.

Table 1 Features of the study population in internal datasets presented as a whole and by HFNC outcome

Feature	Overall	HFNC Failure	HFNC Success	*P* Value
Number of Patients	184	68	116	/
Age, y	68 (66, 78)	68 (65, 79)	68 (66, 76)	0.449
SOFA score	4 (3, 4)	5 (4, 7)	4 (3, 4)	< 0.001
Baseline (Time T0)				
HACOR score	5 (4, 6)	6 (4, 7)	4 (3, 6)	< 0.001
ROX index	7.1 (5.3, 8.4)	6.0 (4.6, 6.9)	7.8 (5.9, 9.1)	< 0.001
HR, bpm	94 (82, 102)	95 (89, 104)	94 (86, 102)	0.549
RR, bpm	28 (25, 30)	30 (25, 36)	27 (24, 28)	< 0.001
FiO_2, %	52 (45, 60)	56 (50, 60)	49 (40, 60)	< 0.001
PaO_2, mmHg	65.6 (60.0, 72.0)	64.1 (59.6, 70.2)	66.5 (60.0,73.8)	0.130
PaO_2/FiO_2, mmHg	134.2 (100.4, 158.3)	118.7 (95.8, 140.0)	143.3 (118.4, 167.8)	< 0.001
SpO_2, %	93 (90, 95)	92 (90, 95)	92 (91, 95)	0.235
$PaCO_2$, mmHg	32.9 (31.2, 34.5)	32.1 (30.4, 34.3)	33.4 (31.6, 34.5)	0.028
SpO_2/FiO_2	188 (150, 211)	170 (148, 186)	199 (158, 235)	< 0.001
2h after HFNC (Time T1)				
HACOR score	4.0 (3.0, 5.0)	5 (4, 6)	3 (2, 4)	< 0.001
ROX index	8.8 (5.8, 11.1)	5.6 (4.8, 6.3)	10.7 (8.3, 13.1)	< 0.001
HR, bpm	92 (80, 102)	95 (86, 103)	90 (80, 100)	0.026
RR, bpm	24 (20, 26)	28 (26, 30)	21 (19, 24)	< 0.001
FiO_2, %	51 (40, 63)	63 (55, 70)	46 (35, 51)	< 0.001
PaO_2, mmHg	66.4 (60.7, 71.1)	67.2 (60.4, 72.7)	66.0 (61.2, 70.0)	0.152
PaO_2/FiO_2, mmHg	141.9 (110.8, 165.8)	114.9 (96.5, 131.0)	157.7 (129.5, 192.3)	< 0.001
SpO_2, %	94 (93, 95)	94 (93, 95)	94 (92, 95)	0.958
$PaCO_2$, mmHg	34.4 (32.3, 36.5)	32.6 (30.8, 35.0)	35.4 (33.2, 36.8)	< 0.001
SpO_2/FiO_2	199 (148, 234)	155 (134, 171)	224 (182, 271)	< 0.001

T0 represents the time point when patients' characteristics were collected at the initiation of HFNC therapy. T1 represents the time point when patients' characteristics were collected 2 h after the initiation of HFNC therapy. The HACOR score is a clinical scoring system used to assess the risk of HFNC failure, incorporating five parameters: Heart rate, Acidosis (pH), Consciousness level, Oxygenation (PaO_2/FiO_2 ratio), and Respiratory rate. The SOFA (Sequential Organ Failure Assessment) score is a clinical tool used to evaluate the extent of organ dysfunction or failure, based on six systems: respiratory, cardiovascular, hepatic, coagulation, renal, and neurological functions

Performance metrics excluding patients with SpO2 > 97%: Some studies [42, 43] have suggested that clinical indices involving SpO_2 may exhibit reduced accuracy when SpO_2 exceeds 97%, due to the flat portion of the oxyhemoglobin dissociation curve, an increased physiological overtreatment rate, and an increased potential for measurement error. To explore this, we excluded patients with SpO2 > 97% at any time point, resulting in an exclusion of 3 patients from the internal dataset and 122 patients from the external dataset. As shown in Additional file: Tables S5 and S6, Fig. S6, we observed an increased performance in external validation for SpO_2 related indices, with the accuracy of the ROX-HR index increasing from 67 to 70%. A comparison of the permutation importance plots (Fig. 4a and Additional file: Fig. S7b) reveals an increased significance of SpO_2-related measurements. The developed ML model also demonstrated performance improvements, with the accuracy of the model improving from 73 to 75%. The ML model again consistently outperformed all clinical indices in terms of overall accuracy and discrimination ability in this cohort (Additional file: Fig. S6a).

Predictive performance metrics including arterial blood gas (ABG) measurements: Among the developed ML models, the SVM again performed best in both the internal and external validation when ABG measurements were included in the training dataset (Additional file: Table S2, Fig. S5). In internal validation using measurements from two time points, the model achieved 92% accuracy, 94% sensitivity, 92% specificity, and 0.96 AUC in LOOCV. Additionally, in nested five-fold cross-validation with 95% confidence intervals, the model had an accuracy of 89% (95% CI: 82–97%), sensitivity of 89% (95% CI: 74–100%), specificity of 88% (95% CI: 77–100%), and an AUC of 0.89 (95% CI: 0.80–0.98). In external validation, on the MIMIC-IV/eICU datasets, the SVM model achieved an accuracy of 83%, a sensitivity of 84%, a specificity of 82%, and an AUC of 0.82 (see Fig. 3b). The DCA

Table 2 Comparative performance of machine learning models and clinical indices on internal and external datasets without using ABG measurements

Time Point	Validation Methods	Accuracy	Sensitivity	Specificity	AUC
SVM ML Model Without Using ABG Measurements					
T0	LOOCV	69%	35%	89%	0.62
	RNFCV	67% [53%, 74%]	46% [27%, 66%]	82% [79%, 100%]	0.66 [0.50, 0.74]
	External Validation	67%	63%	68%	0.65
T1	LOOCV	88%	87%	89%	0.89
	RNFCV	86% [78%, 96%]	85% [73%, 100%]	87% [69%, 94%]	0.87 [0.76, 0.96]
	External Validation	72%	68%	74%	0.73
T0 + T1	LOOCV	90%	88%	91%	0.93
	RNFCV	87% [76%, 95%]	88% [71%, 100%]	89% [76%, 97%]	0.87 [0.78, 0.97]
	External Validation	73%	73%	73%	0.79
Clinical Scores and Indices Without Using ABG Measurements					
ROX (T0)	LOOCV	64%	31%	77%	0.61
	RNFCV	65% [31%, 77%]	44% [21%, 64%]	79% [65%, 100%]	0.64 [0.48, 0.76]
	External Validation	61%	68%	59%	0.68
ROX (T1)	LOOCV	89%	85%	91%	0.91
	RNFCV	84% [76%, 94%]	82% [64%, 100%]	92% [83%, 100%]	0.85 [0.81, 0.96]
	External Validation	64%	79%	60%	0.74
ROX-HR (T0)	LOOCV	66%	32%	86%	0.68
	RNFCV	66% [54%, 78%]	33% [54%, 78%]	87% [70%, 100%]	0.60 [0.46, 0.73]
	External Validation	67%	62%	69%	0.68
ROX-HR (T1)	LOOCV	84%	82%	85%	0.88
	RNFCV	84% [73%, 95%]	81% [57%, 100%]	87% [70%, 100%]	0.83 [0.71, 0.94]
	External Validation	64%	82%	60%	0.75
RR (T0)	LOOCV	68%	31%	91%	0.63
	RNFCV	68% [57%, 76%]	31% [7% 50%]	91%[78%, 100%]	0.61 [0.50, 0.70]
	External Validation	65%	47%	70%	0.58
RR (T1)	LOOCV	82%	78%	85%	0.87
	RNFCV	84% [72%, 93%]	81% [56%, 100%]	88% [72%, 100%]	0.85 [0.75, 0.95]
	External Validation	63%	61%	64%	0.65
SpO_2/FiO_2 (T0)	LOOCV	60%	28%	79%	0.66
	RNFCV	62% [51%, 73%]	30% [7.1%, 50%]	80% [65%, 100%]	0.55 [0.44, 0.67]
	External Validation	63%	66%	62%	0.66
SpO_2/FiO_2 (T1)	LOOCV	79%	68%	85%	0.85
	RNFCV	78% [68%, 89%]	68% [43%, 86%]	86% [70%, 96%]	0.76 [0.63, 0.88]
	External Validation	59%	86%	51%	0.72
SOFA (before HFNC)	LOOCV	73%	37%	94%	0.52
	RNFCV	72% [62%, 81%]	37% [14%, 57%]	94% [83%, 100%]	0.65 [0.53, 0.78]
	External Validation	56%	42%	66%	0.53

LOOCV: Nested leave-one-out cross-validation. RNFCV: Repeated nested five-fold cross-validation. Internal validation was conducted using various nested cross-validation methods, including LOOV and RNFCV on the training set. External validation was performed on a separate external dataset. The internal dataset consists of 184 patients (116 successes and 68 failures), while the external dataset includes 567 patients (440 successes and 127 failures). T0 represents the time point when patients' characteristics were collected at the initiation of HFNC therapy. T1 represents the time point when patients' characteristics were collected 2 h after the initiation of HFNC therapy. T0 + T1 represents the combined measurements taken at both T0 (at HFNC initiation) and T1 (2h after HFNC initiation). Numbers in the square brackets indicate the 95% confidence intervals, ranging from the 2.5th to the 97.5th percentile

(Fig. 3d) and calibration curve (Additional file: Fig. S3a) again confirm that the model is well calibrated and provides greater net benefit compared to simpler strategies.

Among the currently available clinical indices/scores in our datasets, mROX (which uses PaO_2) evaluated at time T1 achieved the highest predictive performance in external validation, with an accuracy of 70%, sensitivity of 63%, specificity of 74% and an AUC of 0.65. The Brier score for mROX is higher than that of the SVM model, indicating poorer calibration, particularly within intermediate- and

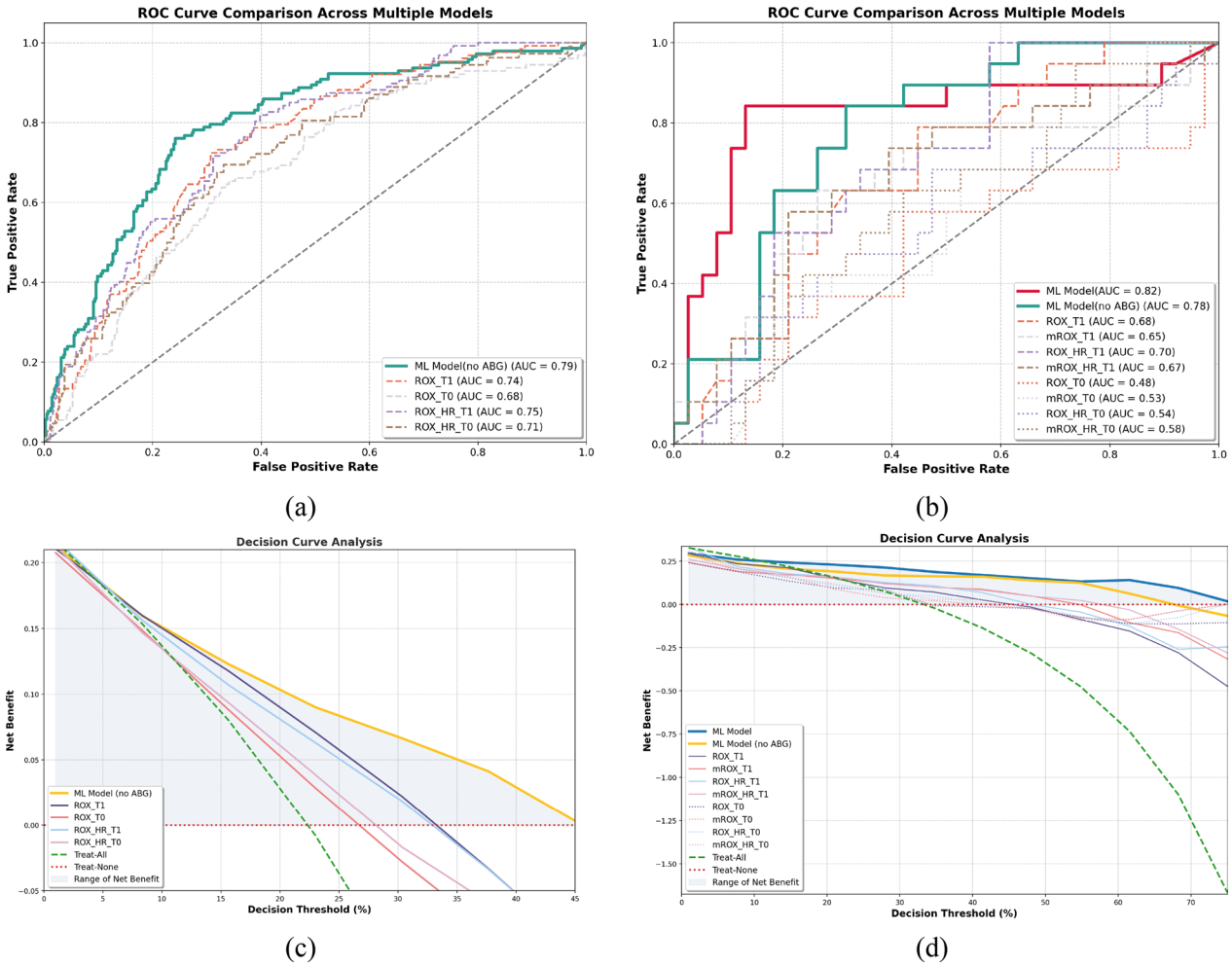

Fig. 3 Graphical comparison of the performance of both machine learning models with the ROX index and its variants, on the external validation cohort. **a, b** The Receiver Operating Characteristic (ROC) curve comparing the performance of various machine learning models and the ROX index, along with its variants, on the external validation cohort. **c, d** Decision curve analysis (DCA) for SVM model and ROX with its variants on the external validation cohort. The curves demonstrate the net benefit across a range of threshold probabilities, comparing the predictive performance of each model against the 'Treat-All' (escalate oxygenation treatment for all patients) and 'Treat-None' (do not escalate oxygenation treatment) strategies. The shaded regions represent the range of net benefit for SVM. Net benefit = (true positives/total patients)—(false positives/total patients) × ω, where ω is the odds at the threshold probability. **Left Panels:** ROC and DCA curves show the performance of a SVM model without using ABG measurements—because such measurements were unavailable in the RENOVATE dataset at matching time points—along with ROX indices. The evaluation was conducted on the combined eICU, MIMIC-IV, and RENOVATE dataset, which includes 567 patients (440 successes and 127 failures). **Right Panels:** ROC and DCA curves comparing the performance of a SVM model with all measurements, a SVM model without using ABG measurements, and ROX indices with variants on the eICU + MIMIC-IV dataset (57 patients: 38 successes, 19 failures)

high-risk patient groups, where the predicted probability of HFNC failure lies within the intermediate to high probability range (see Fig. S6). From the DCA (Fig. 3d), the net benefit for mROX(T1) under different decision thresholds is similar, and only informs meaningful decisions over full intervention or no intervention strategies for thresholds between 30 and 50%. In comparison, the SVM model consistently has a higher net benefit across all thresholds. Although some predictors and indices showed good separation between the success and failure groups in the training set (Additional file: Fig. S9), their distributions exhibited considerable overlap between the two groups, and the optimal cut-off values varied across different datasets (Additional file: Fig. S10). This variability resulted in inconsistent performance across all validation methods for individual indices or predictors.

Predictive performance metrics across timepoint measurements: In both internal and external validation, clinical indices measured at timepoint T1 showed notably higher overall accuracy, sensitivity, specificity,

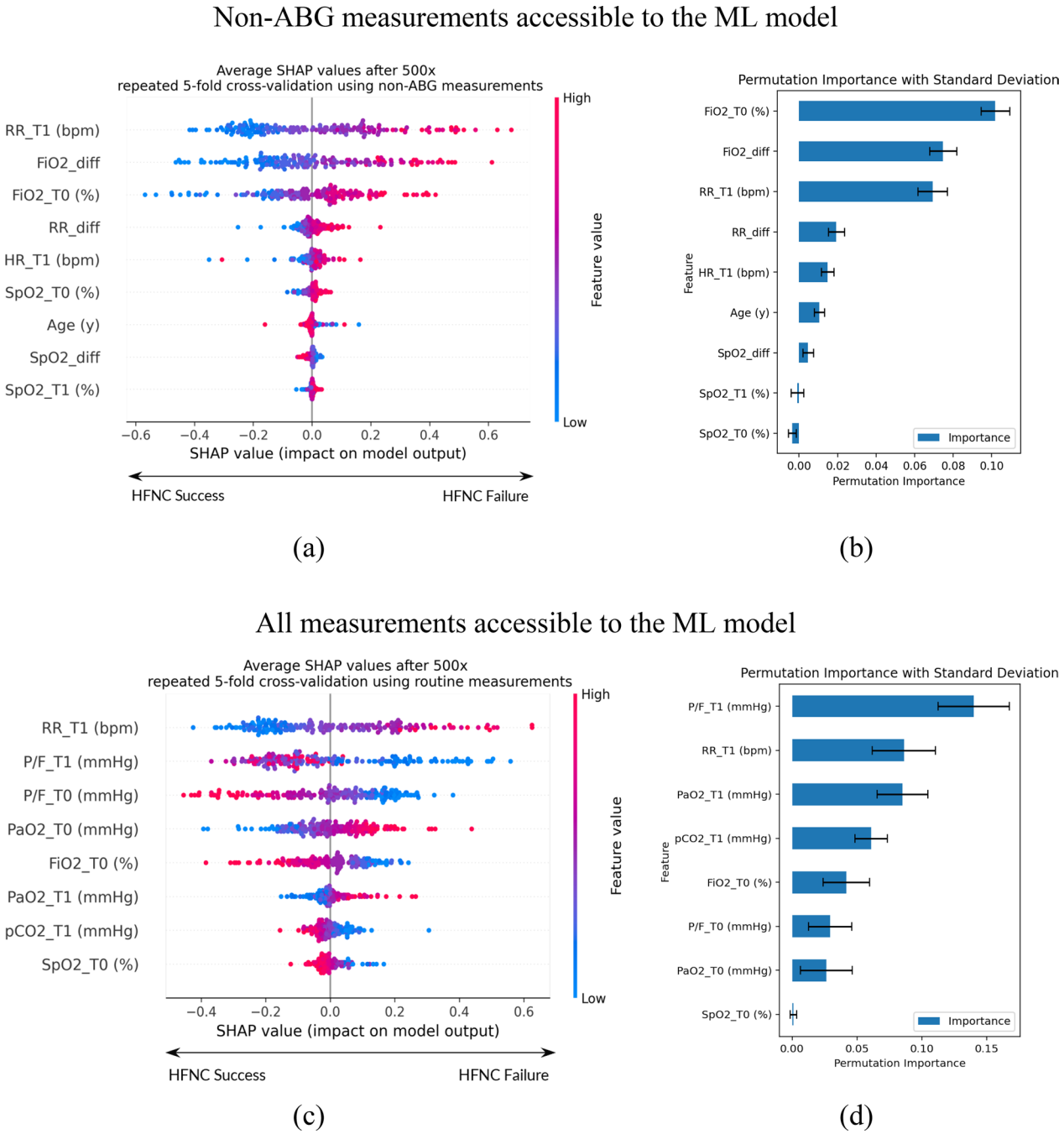

Fig. 4 SVM model explainabilty analysis: SHAP summary plot and permutation importance plot. **a**, **c** Average SHAP values obtained from 500 iterations of repeated fivefold cross-validation on internal dataset. Horizontal axis: The impact of each feature on the model's prediction. Positive SHAP values indicate that the feature contributes to predicting a HFNC failure, while negative SHAP values indicate a contribution to predicting HFNC success. Vertical axis: The list of features used for making a prediction, ordered based on their importance, with the most important features at the top. Dots: Each dot represents a single patient in the dataset. Colour gradient: Indicates the feature value for each observation, where blue represents low feature values and red represents high feature values. **b**, **d** Permutation importance for external validation obtained by SVM model on external dataset. The suffix "_diff" indicates the change in a specific variable between two time points. For instance, "FiO2_diff" represents the change in FiO2 values between T1 and T0 (FiO2_diff = FiO2_T1—FiO2_T0)

and AUC compared to those measured at timepoint T0 (Tables 2 and 3). The differences in sensitivity between clinical indices obtained at the two timepoints are very significant, indicating that clinical indices measured 2 h after HFNC initiation are more reliable in predicting treatment outcome. For the SVM models, using only measurements taken at 2h after HFNC onset for training also provides more reliable predictions compared with using only baseline T0 measurements. When using measurements from both timepoints for training, further significant increases in predictive accuracy were observed, highlighting the ability of the ML models to exploit trends in the measurement data in making predictions. In contrast, when examining the trends of clinical indices (Additional file: Table S3), it can be observed that none of the changes in these indices

Table 3 Comparative performance of various clinical scores and indices using ABG measurements across internal and external datasets

Time Point	Validation Methods	Accuracy	Sensitivity	Specificity	AUC
SVM ML Model Using All Measurements at Different Time Points					
T0	LOOCV	70%	27%	95%	0.65
	RNFCV	69% [60%, 78%]	27% [7%, 50%]	95% [83%, 100%]	0.61 [0.50, 0.72]
	External Validation	67%	47%	76%	0.66
T1	LOOCV	86%	84%	91%	0.95
	RNFCV	87% [78%, 95%]	86% [67%, 100%]	87% [74%, 100%]	0.86 [0.78, 0.96]
	External Validation	74%	90%	66%	0.74
T0+T1	LOOCV	92%	94%	92%	0.96
	RNFCV	89% [82%, 97%]	89% [74%, 100%]	88% [77%, 100%]	0.89 [0.80, 0.98]
	External Validation	83%	84%	82%	0.82
Clinical Scores and Indices Using All Measurements					
mROX (T0)	LOOCV	64%	41%	77%	0.69
	RNFCV	65% [54%, 78%]	43% [21%, 71%]	79% [61%, 96%]	0.61 [0.48, 0.76]
	External Validation	65%	32%	82%	0.53
mROX (T1)	LOOCV	88%	82%	91%	0.90
	RNFCV	88% [78%, 97%]	83% [57%, 100%]	92% [83%, 100%]	0.87 [0.74, 0.96]
	External Validation	70%	63%	74%	0.65
mROX-HR (T0)	LOOCV	66%	37%	84%	0.68
	RNFCV	67% [54%, 76%]	37% [14%, 57%]	85% [65%, 96%]	0.61 [0.49, 0.71]
	External Validation	63%	26%	82%	0.58
mROX-HR (T1)	LOOCV	80%	74%	85%	0.87
	RNFCV	81% [70%, 90%]	74% [50%, 93%]	85% [70%, 96%]	0.80 [0.66, 0.90]
	External Validation	70%	58%	76%	0.67
PaO_2/FiO_2 (T0)	LOOCV	61%	29%	80%	0.66
	RNFCV	62% [51%, 70%]	31% [14%, 50%]	81% [65%, 96%]	0.56 [0.45, 0.66]
	External Validation	47%	32%	55%	0.45
PaO_2/FiO_2 (T1)	LOOCV	74%	59%	83%	0.80
	RNFCV	74% [62%, 87%]	58% [36%, 79%]	84% [70%, 96%]	0.71 [0.57, 0.85]
	External Validation	65%	63%	66%	0.61
HACOR (T0)	LOOCV	71%	29%	94%	0.64
	RNFCV	69% [57%,78%]	31% [7%, 57%]	93% [56%,100%]	0.62 [0.51,0.73]
	External Validation	/	/	/	/
HACOR (T1)	LOOCV	75%	71%	78%	0.72
	RNFCV	74% [62%, 84%]	64% [21%, 56%]	80% [61%, 100%]	0.72 [0.59, 0.84]
	External Validation	/	/	/	/

T0 represents the time point when patients' characteristics were collected at the initiation of HFNC therapy. T1 represents the time point when patients' characteristics were collected 2 h after the initiation of HFNC therapy. LOOCV: Nested leave-one-out cross-validation. RNFCV: Repeated nested five-fold cross-validation. Numbers in the square brackets indicate the 95% confidence intervals, ranging from the 2.5th to the 97.5th percentile. Since ABG measurements were unavailable in the RENOVATE dataset at matching time points, the external dataset used for evaluating the model or indices with ABG measurements was based on the eICU/MIMIC-IV dataset, including 57 patients (38 successes and 19 failures)

exhibit strong predictive ability, with their predictive performance being lower than that of the indices measured only at T1.

Model interpretability and relative importance of patient measurements: Model interpretability was examined through SHAP summary plots and permutation importance. The SHAP summary dot plot (Fig. 4a, c) presents the contribution of individual measurement to the HFNC outcome predicted by the SVM model during repeated five-fold CV on the training set, and the permutation importance (Fig. 4b, d) was calculated by randomly shuffling the values of each measurement in external validation, illustrating the relative impact of each measurement on model performance. Figure 4a, b also shows the SHAP values and permutation importance of the model when ABG measurements are not used. When all measurements are included, the SHAP plot shows that among these measurements, RR and PaO_2/FiO_2 ratio at 2h after HFNC initiation were the strongest factors in predicting HFNC failure, with higher values of RR at T1 or lower values of PaO_2/FiO_2 at T1 associated with an increased risk of HFNC failure. The permutation importance plot generated from the external validation dataset also showed that measurements taken at T1 are ranked highest, with RR and PaO_2/FiO_2 ratio at T1 again the most critical measurements for the model's predictive accuracy. Additional file: Fig. S8 shows that there are strong correlations between RR and ABG measurements such as PaO_2 and $PaCO_2$, which exhibit a high degree of interdependence when making predictions. Regardless of whether ABG data is available for the model, SpO_2 was always of lesser importance. The PaO_2/FiO_2 ratio holds significant value in the model when ABG measurements are available, whereas SpO_2 does not. These results suggest that the predictive accuracy of ROX may be primarily determined by RR and FiO_2, as RR at T1 and the change in FiO_2 between T0 and T1 ranked highest in importance in the model when ABG measurements are not used. In the cohort where patients with SpO2 greater than 97% were excluded, the strong dependency of ROX on RR and FiO_2 is partially alleviated due to the enhanced discriminatory power of SpO_2 within this range. However, despite this improvement, the feature importance of SpO_2 in this smaller cohort remains low (Additional file: Fig. S7). Finally, as shown in Fig. 4c, lower values of $PaCO_2$ (blue dots) at T1 consistently contribute to the model predicting HFNC failure (positive x-axis values), whereas higher values of $PaCO_2$ (red dots) consistently contribute to the model predicting HFNC success (negative x-axis values), suggesting a role for high respiratory drive in determining HFNC failure.

Discussion

In this study, we evaluated the predictive performance of machine learning models for predicting HFNC outcomes and compared the results to the performance of currently used clinical indices. All models were rigorously evaluated using nested cross-validation methods, and further validated on external datasets from multiple different sources. To the best of our knowledge, this study is the first to develop an externally validated ML model for prediction of HFNC outcome using measurements taken only in the very early stages of treatment. Two previous studies have attempted to develop ML models to predict HFNC outcome [32, 33], however these studies performed no external validation, and used data from unspecified time points over 24- and 8-h time windows to make predictions.

The accuracy of the SVM model was superior to that of all clinical indices that have recently been proposed as potential predictors of HFNC outcomes. Among these, the ROX index and its modified versions, mROX, ROX-HR, and mROX-HR were the most accurate predictors of HFNC outcome, as established in multiple previous studies [10, 14, 34–38]. However, external validation is not common in these studies, and in some cases where only internal datasets are available, a separate validation cohort was not utilized. This limitation is compounded by significant heterogeneity across the studies, including variations in the criteria for initiating HFNC therapy between centres, onset of intubation, HFNC initiation flow settings (ranging from 10 to 60 L/min), and definition of HFNC failure, which span from 48 h to 28 days post-initiation. These factors may help explain the substantial variation in the performance of the ROX index when evaluated at the same timepoint across different studies. Moreover, some previous studies [37] have reported low discrimination and poor calibration of ROX at 2 h, with an AUC of 0.57. In [37], for example, approximately 40%, 57%, and 53% of patients in the identified low-, intermediate-, and high-risk groups, respectively, experienced treatment failure, indicating poor stratification of ROX indices, which aligns with the results found here. Although recently proposed alternative indices, such as VOX [34], esophageal pressure swing (ΔP_{es}) [18] and nasal pressure swing (ΔP_{nose}) [19], have shown good predictive performance, these indices are not widely available in clinical practice.

Our results shed new light on the relative importance in predicting HFNC outcome of those patient measurements which are currently available. Crucially, the PaO_2/FiO_2 ratio holds significant predictive value in the ML model when ABG measurements are available, whereas SpO_2/FiO_2 does not. The notable increase in predictive accuracy provided by ML models which have access to

ABG measurements suggests that there could be significant benefits to taking arterial blood gasses just before, and in the early stages of HFNC therapy, particularly in patients with severe AHRF whose peripheral oxygen saturation could be influenced by other factors (i.e. hemodynamics). A second interesting point relates to the ability of the ML models to exploit trends in the measurement data over this initial time period in making predictions. This contrasts with current clinical indices, whose changes over the first two hours of treatment did not exhibit strong predictive power (Additional file: Tables S3, and S4).

Our study has several strengths. First, the use of three different datasets from North and South America for external validation instead of a specific subset of patients from one institution increases confidence in the generalizability of the model. Second, both internal and external cross-validations were performed, and the developed models were compared against optimized thresholds of standard clinical indices using the same validation methods, in order to fairly compare their relative performance and robustness. Third, a more reliable and unbiased nested repeated cross-validation approach was employed for internal performance evaluation. Unlike conventional cross-validation, which relies on a single run for model selection and performs feature selection prior to validation, nested repeated cross-validation mitigates the risk of overfitting, reduces performance overestimation, and provides more stable and reliable performance estimates with lower variance, particularly for small datasets [39, 40]. Fourth, all data were collected at or before the time of HFNC initiation and at 2 h post-initiation, providing an accurate test of the feasibility of early prediction of HFNC failure. Fifth, advanced feature engineering techniques were integrated into the nested cross-validation framework, including a genetic feature search algorithm, which facilitates an extensive exploration of a broad feature space and captures non-linear interactions between features (see Fig. 2). Finally, the model development followed all steps outlined in the guidelines proposed in the recent publication: Developing Clinical Prediction Models: A Step-by-step Guide [41].

Some limitations must also be carefully considered before incorporation of the proposed ML models in clinical practice. First, although the granularity of the measurements available in the analysed datasets is high, the total number of patients available in our datasets is still smaller than what would typically be used for the training and validation of ML models. Second, the requirement for measurements at consistent time points restricted the external validation cohort for the ML models using ABG measurements to the 57 patients from the eICU and MIMIC-IV databases, who had all measurements available at the required time points. Third, we did not assess the predictive power of some physiological variables (e.g. inspiratory effort and tidal volumes), as they are not routinely measured during HFNC therapy and are absent from data on patients treated with HFNC in the external datasets. Last, it must be emphasised that decisions regarding modification or escalation of non-invasive respiratory support are inherently extremely complex—while models of the type considered here can provide very useful information, they are no substitute for clinical judgement.

Conclusions

In this study, we developed and tested machine learning models to predict the risk of HFNC treatment failure in patients with AHRF, using only measurements made in the first two hours of the patients' course of treatment. Models based on the support vector machine algorithm demonstrated improved predictive accuracy, discrimination and calibration compared to optimized thresholds calculated for currently used clinical indices. While real-time decision support tools based on machine learning models should never be seen as a replacement for careful clinical judgement, our results suggest that they could provide valuable real-time assistance to clinicians in identifying patients who are at higher risk of failing HFNC, allowing their treatment to be adjusted or escalated in a timely fashion.

Abbreviations

HFNC	High-flow nasal cannula
ML	Machine learning
AHRF	Acute hypoxemic respiratory failure
ICU	Intensive care unit
ML	Machine learning
SHAP	Shapely additive explanation
ROX	Ratio of oxygen saturation as measured by pulse oximetry/oxygen fraction index
RR	Respiratory rate
HR	Heart rate
SpO_2	Saturation of pulse oxygen
PaO_2/FiO_2	The ratio of partial pressure of oxygen in arterial blood to the fraction of inspiratory oxygen concentration
VOX	The ratio of SpO_2/FiO_2 over tidal volume
ABG	Arterial blood gas
CV	Cross validation
LOOCV	Nested leave-one-out cross-validation
RNFCV	Repeated nested five-fold cross-validation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13054-025-05336-4.

Additional file1 (DOCX 13649 KB)

Acknowledgements

Not applicable.

Author contributions
DGB, SS and HY designed the study. HY was the principal modeler and created the machine learning models, with contributions from SS, RT, EC, JGL and DGB. RT, EC, LL, LK-D, IM, and AC provided individual patient data and assisted in development of the models. All authors analyzed and interpreted new data and results in the context of existing literature and clinical practice. DGB and HY wrote the manuscript. All authors reviewed and revised the manuscript.

Funding
This research was supported by the UKRI Engineering and Physical Sciences Research Council (Ref. EP/W000490/1) and The Royal Academy of Engineering (Ref. RF2122-21–258).

Availability of data and material
Patient data for internal validation and code used for the machine learning model are freely available on request by bona fide researchers for specified scientific purposes from the corresponding author. Data for external validation are publicly accessible, including deidentified patient data from the Medical Information Mart for Intensive Care IV (MIMIC-IV) v2.2 database at https://physionet.org/content/mimiciv/2.2/ and from the eICU Collaborative Research Database (eICU) at https://eicu-crd.mit.edu/. The RENOVATE dataset used for external validation is available upon request to bona fide researchers for specific scientific purposes, subject to approval by the RENOVATE investigators. A web-based application based on the model is currently under development.

Declarations

Ethics approval and consent to participate
Not Applicable.

Consent for publication
Not Applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Engineering, University of Warwick, Coventry CV4 7AL, UK. [2]Department of Medical and Surgical Sciences of Adult and Mother-Child SMECHI-MAI, University of Modena Reggio-Emilia, Via del Pozzo, 71, 41124 Modena, Italy. [3]Respiratory Diseases Unit, University Hospital of Modena Policlinico, Modena, Italy. [4]Anaesthesia and Intensive Care Medicine, Galway University Hospitals, Galway, Ireland. [5]School of Medicine, University of Galway, Galway, Ireland. [6]Intensive Care Unit, Hospital Morales Meseguer, Murcia, Spain. [7]Hcor Research Institute, Hcor Hospital, Rua Desembargador Eliseu Guilherme, 200 Paraíso, São Paulo 04004-030, Brazil.

Received: 10 December 2024 Accepted: 22 February 2025
Published online: 07 March 2025

References

1. Frat JP, Thille AW, Mercat A, Girault C, Ragot S, Perbet S, et al. High-flow oxygen through nasal cannula in acute hypoxemic respiratory failure. N Engl J Med. 2015;372:2185–96.
2. Frat JP, Ragot S, Girault C, Perbet S, Prat G, Boulain T, et al. REVA network effect of non-invasive oxygenation strategies in immunocompromised patients with severe acute respiratory failure: a post-hoc analysis of a randomised trial. Lancet Respir Med. 2016;4:646–52.
3. Oczkowski S, Ergan B, Bos L, et al. ERS clinical practice guidelines: high-fow nasal cannula in acute respiratory failure. Eur Respir J. 2022;59(4):2101574.
4. Evans L, Rhodes A, Alhazzani W, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock. Intensive Care Med. 2021;47(11):1181–247.
5. Spinelli E, Mauri T, Beitler JR, Pesenti A, Brodie D. Respiratory drive in the acute respiratory distress syndrome: pathophysiology, monitoring, and therapeutic interventions. Int Care Med. 2020;46(4):606–18.
6. Kang BJ, Koh Y, Lim CM, Huh JW, Baek S, Han M, et al. Failure of high flow nasal cannula therapy may delay intubation and increase mortality. Int Care Med. 2015;41:623–32.
7. Yasuda H, Okano H, Mayumi T, Nakane M, Shime N. Association of non-invasive respiratory support with mortality and intubation rates in acute respiratory failure: a systematic review and network meta-analysis. J Int Care. 2021;9(1):32.
8. Grasselli G, Calfee CS, Camporota L, et al. ESICM guidelines on acute respiratory distress syndrome: definition, phenotyping and respiratory support strategies. Int Care Med. 2023;49(7):727–59.
9. Liu T, Zhao Q, Du B. Effects of high-flow oxygen therapy on patients with hypoxemia after extubation and predictors of reintubation: a retrospective study based on the MIMIC-IV database. BMC Pulm Med. 2021;21(1):160.
10. Roca O, Caralt B, Messika J, Samper M, Sztrymf B, Hernandez G, et al. An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy. Am J Respir Crit Care Med. 2019;199:1368–76.
11. Goh KJ, Chai HZ, Ong TH, Sewa DW, Phua GC, Tan QL. Early prediction of high flow nasal cannula therapy outcomes using a modifed ROX index incorporating heart rate. J Int Care. 2020;8:41.
12. Karim HMR, Esquinas AM. Success or failure of high-fow nasal oxygen therapy: the ROX Index is good, but a modifed ROX index may be better. Am J Respir Crit Care Med. 2019;200(1):116–7.
13. Arunachala S, Parthasarathi A, Basavaraj CK, Kaleem Ullah M, Chandran S, Venkataraman H, Vishwanath P, Ganguly K, Upadhyay S, Mahesh PA. The validity of the ROX Index and APACHE II in predicting early, late, and non-responses to non-invasive ventilation in patients with COVID-19 in a low-resource setting. Viruses. 2023;15(11):2231.
14. Mellado-Artigas R, Mujica LE, Ruiz ML, et al. Predictors of failure with high-flow nasal oxygen therapy in COVID-19 patients with acute respiratory failure: a multicenter observational study. J Int Care. 2021;9:23.
15. Valencia CF, Lucero OD, Castro OC, Sanko AA, Olejua PA. Comparison of ROX and HACOR scales to predict high-flow nasal cannula failure in patients with SARS-CoV-2 pneumonia. Sci Rep. 2021;11:22559.
16. Yarnell CJ, Johnson A, Dam T, et al. Do thresholds for invasive ventilation in hypoxemic respiratory failure exist? a cohort study. Am J Respir Crit Care Med. 2023;207(3):271–82.
17. Fleuren LM, Thoral P, Shillan D, et al. Machine learning in intensive care medicine: ready for take-off? Int Care Med. 2020;46(7):1486–8.
18. Tonelli R, Fantini R, Bruzzi G, et al. Effect of high flow nasal oxygen on inspiratory effort of patients with acute hypoxic respiratory failure and do not intubate orders. Intern Emerg Med. 2024;19:333–42.
19. Tonelli R, Cortegiani A, Fantini R, et al. Accuracy of nasal pressure swing to predict failure of high-flow nasal oxygen in patients with acute hypoxemic respiratory failure. Am J Respir Crit Care Med. 2023;207(6):787–9.
20. RENOVATE Investigators and the BRICNet Authors. High-flow nasal oxygen vs noninvasive ventilation in patients with acute respiratory failure: The RENOVATE randomized clinical trial. JAMA. 2024. https://doi.org/10.1001/jama.2024.26244.
21. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023. https://doi.org/10.1038/s41597-022-01899-x.
22. Pollard T, Johnson A, Raffa J, et al. The eICU Collaborative research database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178.
23. Potter M, Yıldız A Y, Prabhu N M, et al (2024) Distributed $lection. arXiv. https://doi.org/10.48550/arXiv.2401.10846
24. Liaw R, Liang E, Nishihara R, et al (2018) Tune: a research platform for distributed model selection and training. arXiv. https://doi.org/10.48550/arXiv.1807.05118
25. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit. 2015;48:2839–46.
26. Xu Y, Goodacre R. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. J Anal Test. 2018;2:249–62.
27. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? J Am Stat Assoc. 2024;119(546):1434–45.
28. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14:1–13.

29. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA. 2015;313(4):409–10.
30. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Comput Methods Programs Biomed. 2022;214:106584.
31. Mi X, Zou B, Zou F, et al. Permutation-based identification of important biomarkers for complex diseases via machine learning models. Nat Commun. 2021;12:3008.
32. Wang Z, Chao Y, Xu M, et al. Machine learning prediction of the failure of high-flow nasal oxygen therapy in patients with acute respiratory failure. Sci Rep. 2024;14:1825.
33. Hendriks M (2023) An AI algorithm to predict intubations in ICU patients-Predicting intubation to aid medical personnel in the decision to switch from High Flow Nasal Oxygen therapy to intubation. Master thesis, Delft University of Technology. TU Delft Repository. https://resolver.tudelft.nl/uuid:1d7f9cb8-6ccd-4895-872d-2bf51d5686e0
34. Chen D, Heunks L, Pan C, et al. A novel index to predict the failure of high-flow nasal cannula in patients with acute hypoxemic respiratory failure: a pilot study. Am J Respir Crit Care Med. 2022;206(7):910–3.
35. Li Z, Chen C, Tan Z, et al. Prediction of high-flow nasal cannula outcomes at the early phase using the modified respiratory rate oxygenation index. BMC Pulm Med. 2022;22:227.
36. Ferrer S, Sancho J, Bocigas I, et al. ROX index as predictor of high flow nasal cannula therapy success in acute respiratory failure due to SARS-CoV-2. Respir Med. 2021;189:106638.
37. Okano H, Yamamoto R, Iwasaki Y, et al. External validation of the HACOR score and ROX index for predicting treatment failure in patients with coronavirus disease 2019 pneumonia managed on high-flow nasal cannula therapy: a multicenter retrospective observational study in Japan. J Int Care. 2024;12:7.
38. Praphruetkit N, Boonchana N, Monsomboon A, et al. ROX index versus HACOR scale in predicting success and failure of high-flow nasal cannula in the emergency department for patients with acute hypoxemic respiratory failure: a prospective observational study. Int J Emerg Med. 2023;16:3.
39. Krstajic D, Buturovic LJ, Leahy DE, et al. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6:10.
40. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinform. 2006;7:91.
41. Efthimiou O, Seo M, Chalkou K, et al. Developing clinical prediction models: a step-by-step guide. BMJ. 2024. https://doi.org/10.1136/bmj-2023-078276.
42. Wick KD, Matthay MA, Ware LB. Pulse oximetry for the diagnosis and management of acute respiratory distress syndrome. Lancet Respir Med. 2022;10(11):1086–98.
43. van den Boom W, Hoy M, Sankaran J, Liu M, Chahed H, Feng M, See KC. The search for optimal oxygen saturation targets in critically ill patients: observational data from large ICU databases. Chest. 2020;157(3):566–73.

Publisher's Note