Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early prediction of weaning trajectories and outcomes with machine learning in patients receiving Invasive Mechanical Ventilation in the WEAN SAFE cohort

Nadeem Iqbal Kajla
Ollscoil na Gaillimhe – University of Galway
Reginald Caldecott
Ollscoil na Gaillimhe – University of Galway
Emma Urquhart
Ollscoil na Gaillimhe – University of Galway
Bairbre McNicholas
Ollscoil na Gaillimhe – University of Galway
Taï Pham
Bicêtre Hospital
Leo Heunks
Radboud University Medical Center
Giacomo Bellani
University of Trento
Laurent Brochard
Unity Health Toronto
John Laffey
jlaffey@universityofgalway.ie

Ollscoil na Gaillimhe – University of Galway
Michael Madden
Ollscoil na Gaillimhe – University of Galway

Research Article

Keywords: clinical prediction model, weaning, explainable AI, tabular foundation model, TabPFN, AI in ICU

Posted Date: July 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10018593/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Identification of patients with specific weaning trajectories early in weaning from invasive mechanical ventilation (MV) may improve outcomes. Our primary objective was to examine the potential for machine learning models to predict difficult and/or prolonged weaning, and secondly to predict successful first extubation, and to identify contributing factors, at the time of the first separation attempt.

Methods

We analysed 3,643 patients from the WEAN SAFE prospective multinational cohort who underwent at least one separation attempt from invasive MV. Eight machine learning models were compared across two prediction tasks: (1) easy (≤ 24 h) versus difficult/failed (≥ 2 days) weaning and (2) successful first extubation attempt. Explainability analyses were used to identify the most influential features.

Results

For prediction of easy versus difficult/prolonged weaning, the tabular foundation models TabICL and TabPFN models achieved the highest discrimination (ROC-AUC 0.834 and 0.830 respectively), outperforming the other models tested. For prediction of successful first extubation, TabPFN and TabICL again achieved the highest discrimination (ROC-AUC 0.743 and 0.733 respectively). Explainability analysis identified a consistent, clinically coherent cluster of features (day of the first SA, agitation burden, organ dysfunction severity, frailty, sedation state, and respiratory physiology) for both predictions.

Conclusions

Machine learning models achieved good discrimination for distinguishing weaning trajectories early in weaning from invasive MV. Key features driving these predictions are clinically meaningful and potentially modifiable at the time the patient first meets separation criteria.

Trial registration: ClinicalTrials.gov, NCT03255109.

Introduction

The successful weaning of a patient from invasive mechanical ventilation (MV) is a critical step in surviving a critical illness [1]. While most patients wean without difficulty, many endure a prolonged and complicated process. Across multiple cohorts, the weaning phase accounts for roughly 40% of total ventilation duration, concentrating clinical time and resources in this phase of the ICU stay [1–3]. Weaning failure, defined in the WEAN SAFE primary paper as reintubation within 7 days of extubation [1], occurs in 10–20% of patients and is consistently associated with longer ventilation, longer ICU and hospital stays, and mortality rates approaching 25–50% [3]. In patients who ultimately fail to wean from

invasive MV, mortality is extremely high, and weaning failure is in effect the major determinant of ICU survival in this cohort [1, 3].

Identifying patients with a high likelihood of first extubation success, and those at higher risk for difficult or failed weaning, as early as possible in the weaning process may allow targeted interventions to improve outcomes. However, early prediction of weaning success at the individual patient level has traditionally proven difficult. The rapid shallow breathing index, which is the most widely studied predictor, showed high accuracy in its original derivation cohort [4], but subsequent studies and systematic reviews have demonstrated only modest pooled discrimination across broader populations [5, 6]. Spontaneous breathing trials (SBTs) are considered by many to be the 'gold standard' to determine readiness for extubation [7], but variability in the clinical use of SBTs across ICUs and differences in the type of SBT used [8, 9] has led to variability in their performance.

The aim of this work is to examine the potential for machine learning to identify patients at risk for difficult and failed weaning at the time of the first separation attempt. The WEAN SAFE dataset identified patients at an early stage of the weaning process, namely at the point of the first separation attempt [1, 10], a time point at which targeted interventions could be implemented to improve outcomes. In this secondary WEAN SAFE analysis, we address two clinically relevant prediction tasks: (1) easy versus difficult/failed weaning; and (2) successful first extubation, among patients weaning from invasive MV. An additional key aim os to identify features used by the models that may be potentially modifiable at the time of the first separation attempt.

Methods and Materials

Study Design and Setting

This is a sub-study of the WEAN SAFE study, an international, multicentre, prospective cohort study of patients undergoing invasive or non-invasive ventilation, conducted during four consecutive weeks between October 2017 and June 2018 in a convenience sample of 481 intensive care units (ICUs) from 50 countries, across five continents, recruiting 5,869 patients who required at least two days of invasive MV [1]. The study, jointly supported by the European Society of Intensive Care Medicine and the European Respiratory Society, was endorsed by multiple national societies/networks (**Appendix 1**). National coordinators and site investigators (Appendix 1) were responsible for obtaining ethics committee approval and ensuring data integrity and validity.

Participants and Data Collection

All patients admitted to a participating ICU aged > 16 years, receiving invasive MV two calendar days after intubation, and providing informed consent (where required), were included in the WEAN SAFE study. The study population for this analysis consisted of all patients undergoing at least one separation attempt (Fig. 1). Patients transferred to other facilities before successful weaning were deemed lost to

follow-up and their ICU and hospital outcomes were not collected. All data were recorded for each patient at the same time each day within participating ICUs, normally as close as possible to 10 a.m.

Data Definitions

Our data definitions have been previously reported [1]. Briefly, initiation of weaning from invasive MV is defined as the time at which the first separation attempt (SA) from the ventilator was performed [10]. A separation attempt is defined as a short period of either T-tube trial, low respiratory support (i.e. continuous positive airway pressure of ≤ 5 cm H_2O, or pressure support ventilation with PEEP ≤ 5 cm H_2O and pressure support ≤ 7 cm H_2O), a short period of tracheostomy mask oxygenation, or a spontaneous breathing trial (SBT) as declared by the investigator, or extubation without an SBT. Delayed initiation of weaning is defined as a delay of > 1 day from meeting weaning eligibility criteria (WEC) to the first SA [1]. Increased weaning duration is defined as requiring > 1 day to wean from invasive MV following the first SA. The day of the first SA was defined as the number of days from the start of invasive MV to the first SA. For the easy versus difficult/prolonged weaning task, easy weaning was defined as completion within 1 day of the first SA and difficult or prolonged weaning as 2 or more days; this duration threshold is distinct from the > 1 day used above to denote delayed initiation and increased weaning duration.

Overall weaning success was defined as extubation without death or reintubation within the following 7 days, whichever came first. For tracheotomised patients, weaning success is defined as spontaneous ventilation through tracheostomy without any MV for 7 days, or discharge from the intensive care unit with unassisted breathing, whichever came first. Weaning failure is defined as patients who entered the weaning process but either died, were transferred, or were still invasively ventilated at day 90.

The ‘difficult/failed weaning’ group comprises three types of patients: (1) those successfully extubated only after a longer (≥ 2 days) weaning time; (2) those whose weaning or extubation failed; and (3) those who underwent at least one separation attempt but never proceeded to an extubation attempt.

Duration of invasive MV was calculated as the number of days between the date of intubation and the date of extubation in ICU (or death, if the patient died while receiving invasive MV). Survival was assessed at ICU discharge with follow-up extending up to 28 days, or at hospital discharge with follow-up extending up to 90 days.

Study Outcomes

The primary outcome was the ability of ML models to predict two cohorts, namely patients with difficult/failed (≥ 2 days) weaning, and patients with a successful first extubation attempt, at the time of the first separation attempt. The secondary outcome was to identify the most influential features in explainable ML models, with particular attention to those that are potentially modifiable in clinical practice.

Inclusion and Exclusion Criteria

Inclusion criteria were all patients in the WEAN SAFE dataset who underwent at least one separation attempt. We created the analysis cohort by applying, in order, the following exclusion criteria: (1) death before any separation attempt; (2) transfer before any separation attempt; (3) pre-existing limitation of life-sustaining treatment; (4) terminal extubation decision before the first SA; and (5) tracheostomy before the first extubation attempt.

Analysis Methodology

A more detailed description of the ML methodology is available in the online supplement. All pre-processing steps were performed before model training to obtain a complete, consistent, and fully numeric modelling table. We applied simple, prespecified rules to handle missingness. Eight models were evaluated: logistic regression (reference model), a feed-forward neural network (MLP), three gradient-boosted tree algorithms (LightGBM [11], XGBoost [12], CatBoost [13]), two pretrained tabular foundation models (TabPFN [14], TabICL [15]), and an Explainable Boosting Machine (EBM) [16]. Models were selected to span a range of complexity and interpretability. Full model specifications, hyperparameters, and training protocols are provided in the Online Supplement.

All models were trained and evaluated on the same pre-processed dataset. A stratified hold-out strategy was used: 80% of patients for model development and 20% reserved as a held-out test set, with stratification preserving outcome prevalence in each split. Model performance was evaluated using ROC-AUC, and average precision. Confusion matrices were produced at both a fixed threshold (0.5) and a Youden-derived threshold to characterise sensitivity–specificity trade-offs relevant to clinical decision-making. Global feature importance and SHAP summary plots were generated from the TabPFN model to identify the variables most strongly associated with the predicted outcome.

RESULTS

Study Population

Of 5,869 patients screened, 3,643 met inclusion criteria after exclusions (Fig. 1). Among these, 3,497 (96.0%) had at least one extubation attempt, of whom 3,026 (86.5% of those with an extubation attempt; 83.1% of the overall cohort) had a successful first extubation, 471 (13.5%) failed their first extubation attempt, and 146 (4.0% of the overall cohort) underwent a separation attempt but had no subsequent extubation attempt. Within the successful first-extubation group, 2,659 patients (87.9%) were weaned within 24 hours, 281 (9.3%) required 2–7 days and 86 (2.8%) required more than 7 days. Of the 471 patients who failed their first extubation, 16 were ultimately weaned within 1 day, 139 within 2–7 days and 164 after more than 7 days, while 152 were never successfully weaned.

Prediction of Easy versus Difficult/Failed Weaning

For the easy versus difficult/failed weaning analysis, patients were classified according to the total time from the first SA to successful liberation from invasive mechanical ventilation, irrespective of whether the first extubation attempt was successful. This analysis included all 3,643 patients who underwent at least one SA. Overall, 2,675 patients had an easy wean, comprising 2,659 patients with successful first extubation within ≤ 1 day and 16 patients who failed their first extubation attempt but were ultimately successfully weaned within ≤ 1 day. The remaining 968 patients were classified as having difficult/failed weaning. This group comprised 367 patients with successful first extubation after ≥ 2 days of weaning, 303 patients who failed their first extubation attempt but were ultimately successfully weaned after ≥ 2 days, 152 patients who failed their first extubation attempt and ultimately failed weaning, and 146 patients who underwent a first SA but did not proceed to an extubation attempt (Fig. 1; Table 1). TabICL and TabPFN achieved the highest discrimination (AUC 0.834 and 0.830 respectively; Fig. 2A), with CatBoost (0.814), EBM (0.806), LightGBM (0.802) and XGBoost (0.797) closely behind. Logistic regression (0.708) and the neural network (0.675) were clearly inferior. Precision-recall analysis (Figure e1) mirrored the ROC ranking. At the Youden-J operating threshold (Fig. 2B), TabICL was most sensitive for easy weaning (97%, 519/535) at modest specificity (57.7%); TabPFN was slightly more specific (sensitivity 95.7%, specificity 61.3%). CatBoost reached the highest specificity (70.6%) at lower sensitivity (73.6%). EBM, LightGBM and XGBoost gave intermediate trade-offs; logistic regression and the MLP performed worst at the threshold level.

Model Explainability

The most influential features in the TabPFN SHAP summary were the day of the first SA, agitation days, cumulative paralysis days, PaO_2/FiO_2 at the first SA, and adjusted SOFA-based severity measures, with continuous paralysis, paralysis before weaning eligibility, admission source and frailty also highly ranked (Fig. 3).

Patient-Level Explainability

For each patient, the model generated an individual feature-level contribution score for each characteristic included in the SHAP explanation. Patient 1 was a 26-year-old with low day-3 SOFA, no agitation days, not frail, and a day-3 first SA; the largest positive contributions came from low SOFA, absent agitation and frailty status (Fig. 4A). Patient 2 was a 79-year-old with hypoxaemic respiratory failure, with factors leading to a high likelihood of prolonged wean including day-3 SOFA score of 15, agitation, pH 7.29, frailty, and a day-11 first SA.

Prediction of Successful First Extubation

For the successful first-extubation analysis 3,643 patients were included in the analysis cohort who underwent at least one SA. Among those 3,026 patients had successful first extubation. The other group comprised 617 patients who did not achieve successful first extubation, including 471 patients who failed their first extubation attempt and 146 patients who underwent at least one separation attempt but did not proceed to an extubation attempt (Fig. 1; Table 2). TabPFN achieved the highest ROC-AUC

(0.743), followed by TabICL (0.733); with the other ML models achieving lower ROC-AUC scores (Fig. 5A). Precision-recall analysis (Figure e2) was consistent with the ROC ranking. At the Youden-J operating threshold (Fig. 5b), TabICL was most sensitive for successful first extubation (81.8%, 496/606) at 61 false-positive successes. TabPFN was more balanced (sensitivity 75.7%, specificity 56.9%, 53 false-positive successes, the fewest among the top models). The other models performed less well

Table 2
Demographics and outcome data in patients the study population with first extubation success versus first extubation failure (n = 3,643)

Characteristic	Study Population N = 3,643	First Extubation Success, N = 3,026	First Extubation Failure/No-extubation, N = 617	p-value*
Age (years)	63.0 (50.0, 74.0)	63.0 (50.0, 74.0)	66.0 (52.0, 75.0)	0.006
Frailty Status (CFS Score > 4, %)	746 (20.6%)	596 (19.9%)	150 (24.4%)	0.011
BMI (Kgs/m^2)	25.8 (22.8, 29.6)	25.8 (22.8, 29.6)	25.8 (22.8, 29.5)	0.796
Height (cm)	168.0 (160.0, 175.0)	169.0 (160.0, 175.0)	168.0 (160.0, 175.0)	0.253
Weight (kg)	75.0 (62.0, 86.0)	75.0 (62.0, 87.0)	73.5 (62.0, 85.0)	0.573
ICU admission category				
Medical	2,465 (67.7%)	2,031 (67.1%)	434 (70.3%)	0.119
Planned Surgery	323 (8.9%)	281 (9.3%)	42 (6.8%)	0.048
Trauma	299 (8.2%)	249 (8.2%)	50 (8.1%)	0.918
Urgent Surgery	556 (15.3%)	465 (15.4%)	91 (14.7%)	0.697
Cause(s) for ICU admission				
Hypoxemic respiratory failure	1,214 (33.3%)	970 (32.1%)	244 (39.5%)	< 0.001
Sepsis	840 (23.1%)	691 (22.8%)	149 (24.1%)	0.480
Hypercapnic respiratory failure	529 (14.5%)	432 (14.3%)	97 (15.7%)	0.353
Non-traumatic neurologic event	481 (13.2%)	390 (12.9%)	91 (14.7%)	0.213
Emergency surgery	499 (13.7%)	414 (13.7%)	85 (13.8%)	0.950
Airway protection	434 (11.9%)	358 (11.8%)	76 (12.3%)	0.734

The data are n (%), mean ± SD, or median (Q1, Q3) where Q1 and Q3 represent 25th and 75th percentile, respectively. The p-value compares the two groups using Wilcoxon rank-sum test for continuous and Chi-squared test for categorical variables. For ICU admission category, a single global p-value (< 0.01) indicates the overall distribution of categories differs between males and females.

Characteristic	Study Population N = 3,643	First Extubation Success, N = 3,026	First Extubation Failure/No-extubation, N = 617	p-value*
Cardiac arrest	295 (8.1%)	240 (7.9%)	55 (8.9%)	0.415
Comorbidities and risk factors				
Respiratory	761 (20.9%)	624 (20.6%)	137 (22.2%)	0.378
Cardiovascular	519 (14.2%)	437 (14.4%)	82 (13.3%)	0.456
Liver	172 (4.7%)	143 (4.7%)	29 (4.7%)	0.978
Kidney	391 (10.7%)	305 (10.1%)	86 (13.9%)	0.005
Neuromuscular	88 (2.4%)	71 (2.3%)	17 (2.8%)	0.547
Immune Dysfunction	516 (14.2%)	415 (13.7%)	101 (16.4%)	0.085
Diabetes	794 (21.8%)	656 (21.7%)	138 (22.4%)	0.706
Clinical Outcomes				
Total duration of invasive MV, days	6.0 (4.0, 10.0)	5.0 (4.0, 8.0)	14.0 (10.0, 20.0)	< 0.001
Length of ICU stay, days	10.0 (6.0, 15.0)	9.0 (6.0, 13.0)	18.0 (12.0, 26.0)	< 0.001
Length of hospital stay, days	21.0 (13.0, 36.0)	20.0 (13.0, 33.0)	28.0 (17.0, 47.0)	< 0.001
Limitation of life-sustaining interventions	521 (14.3%)	364 (12.0%)	157 (25.4%)	< 0.001
ICU mortality	450 (12.4%)	234 (7.7%)	216 (35.0%)	< 0.001
Hospital mortality	726 (20.0%)	459 (15.2%)	267 (43.3%)	< 0.001

The data are n (%), mean ± SD, or median (Q1, Q3) where Q1 and Q3 represent 25th and 75th percentile, respectively. The p-value compares the two groups using Wilcoxon rank-sum test for continuous and Chi-squared test for categorical variables. For ICU admission category, a single global p-value (< 0.01) indicates the overall distribution of categories differs between males and females.

Model Explainability

For successful first extubation (Figure e3), the day of the first SA was the single most influential predictor, with a later first SA shifting the prediction toward failure. Exemplar patient-Level Explainability analyses for first extubation outcome are presented in Figure e4.

Discussion

Early identification of patients at risk for complicated or prolonged weaning from invasive mechanical ventilation (MV) is clinically important, as it may enable targeted interventions that improve outcomes and optimize resource use. This study evaluated whether machine learning (ML) models could predict both difficult/prolonged weaning trajectories and the likelihood of successful first extubation. In addition, it sought to identify patients at low risk of complications who might benefit from accelerated care pathways.

Overall, ML models demonstrated variable but meaningful discriminatory ability across both prediction tasks. Tabular foundation models, specifically TabPFN and TabICL, consistently performed best, with gradient-boosted trees and explainable boosting machines (EBM) also showing strong performance. Importantly, the models converged on a clinically coherent set of influential features, including timing of the first separation attempt (SA), agitation burden, sedation state, organ dysfunction severity, frailty, and respiratory physiology. Many of these factors are either modifiable or actionable at the bedside, supporting the potential clinical utility of these models.

Prediction of Difficult/Prolonged Weaning

Patients classified as having difficult or prolonged weaning trajectories represented a distinctly higher-risk group. They were older, more frail, and experienced substantially longer durations of ventilation, ICU stay, and hospitalisation, alongside markedly higher mortality rates. These differences highlight the importance of early identification, ideally at the time of the first SA, to allow proactive optimisation strategies. ML models achieved good discrimination in distinguishing easy from difficult/prolonged weaning. TabPFN and TabICL performed best, although other models performing less well. Threshold selection plays a critical role in clinical deployment. Sensitivity-focused thresholds prioritise identifying high-risk patients but may misclassify some as low risk, whereas specificity-focused thresholds reduce false positives at the cost of missing some high-risk cases. Therefore, implementation would require calibration based on institutional priorities and tolerance for risk.

This prediction task is inherently more tractable than first extubation success because it reflects broader illness trajectories rather than a single decision point. The predictive features—such as severity of illness, sedation and agitation levels, neuromuscular blockade exposure, frailty, and respiratory parameters—are already integral to clinical reasoning. Their prominence enhances interpretability and reinforces the relevance of the model outputs in guiding pre-separation optimisation. At the reported levels of discrimination, these models may be suitable to inform clinician decision-making. When comparing two randomly selected patients, one from each group, the model will generally assign a higher probability of difficult weaning to the appropriate individual. However, caution is required, and these models to not substitute for clinician judgement.

Successful First Extubation Attempt

Predicting successful first extubation attempt is highly clinically relevant, as early failure (defined as reintubation within 7 days) occurs in approximately 10–20% of patients and is associated with worse

outcomes, including mortality rates of 25–50% [3]. Identifying patients with a high likelihood of success could support earlier extubation decisions and reduce unnecessary delays. In this study, model performance for this specific task was more modest. TabPFN achieved an area under the receiver operating characteristic curve (ROC-AUC) of 0.743, and TabICL achieved 0.733. While lower than values reported in some retrospective single-centre studies [17–19], these results likely provide a more realistic estimate of performance in a heterogeneous, prospectively collected, multinational cohort anchored to a clearly defined clinical decision point. At the Youden threshold, TabICL demonstrated high sensitivity, correctly identifying a large proportion of patients who would succeed, whereas TabPFN offered a more balanced sensitivity-specificity trade-off with fewer false-positive predictions. As with difficult weaning prediction, threshold selection would influence clinical use, particularly in balancing the risks of premature extubation against unnecessary delays.

Clinically Informative Features

The most influential feature across both prediction tasks was the timing of the first SA, representing the interval from initiation of MV to the first attempt at liberation. This variable likely serves as a proxy for overall illness trajectory and recovery, rather than a directly modifiable parameter. Earlier SA was associated with favourable outcomes, whereas delayed SA correlated with more complex weaning. Neurobehavioral factors, including agitation burden, sedation state, and neuromuscular blockade exposure, were highly influential. These findings highlight the importance of neurological status alongside traditional respiratory parameters. Organ dysfunction severity, as measured by SOFA scores, also contributed significantly, reinforcing the multidimensional nature of weaning readiness.

Respiratory physiology remained central, with dynamic driving pressure and oxygenation (PaO_2/FiO_2) retaining predictive importance. Frailty emerged as a key non-modifiable predictor, in keeping with prior WEAN SAFE analyses [20], and consistent with prior literature linking frailty to poorer weaning outcomes.

These features can be grouped by modifiability. Actively modifiable factors include sedation depth, ventilator settings influencing driving pressure, and aspects of gas exchange. Partially modifiable factors include cumulative agitation and neuromuscular blockade exposure, where ongoing management can still influence outcomes. Non-modifiable or trajectory-based factors, such as frailty and timing of SA, inform baseline risk but cannot be directly altered.

A notable strength of the models is the use of SHAP (Shapley Additive Explanations) values to provide patient-level interpretability. These explanations allow clinicians to understand which factors are driving predictions in individual cases. For example, in a borderline patient, elevated driving pressure and deep sedation may explain a lower predicted probability of success, suggesting specific targets for optimisation prior to extubation. This positions the model as a potential decision-support tool rather than a replacement for clinical judgement.

Potential Clinical Implications

This study advances the existing literature in several important ways. First, unlike several prior studies, which relied on retrospective data and surrogate endpoints [14, 17–19, 21], predictions were anchored to a prospectively recorded first SA, representing a clinically meaningful decision point. Second, our multinational cohort, spanning 50 countries and 481 ICUs, enhances generalisability and reflects real-world variability in practice.

Third, we present the first head-to-head evaluation of tabular foundation models (TabPFN, TabICL) alongside gradient-boosted trees and glass-box models (EBM) on a weaning outcome. TabICL and TabPFN are pre-trained on large synthetic datasets and can be applied to new tasks without task-specific training, offering a potentially scalable approach to clinical prediction [14]. Their strong performance relative to traditional models supports their further exploration in critical care settings. Fourth, the dual-task design, addressing both weaning trajectory and first extubation success, provides a comprehensive evaluation of clinically relevant questions using a unified dataset. This approach allows consistent comparison of feature importance and model behaviour across related but distinct outcomes.

However, these strengths come with trade-offs. The lower discrimination observed for first extubation success reflects the complexity of real-world decision-making and the heterogeneity of the study population. By anchoring predictions to a defined clinical event rather than inferred physiological signals, the study prioritises clinical relevance over maximal predictive performance.

Study Limitations

Several limitations should be acknowledged. Model evaluation was conducted using a held-out subset of the same cohort, and external validation in independent populations is necessary before clinical implementation. Prediction models often experience performance degradation when applied to new settings, particularly across varying clinical practices [22]. The models were developed using structured, high-quality research data, and their performance with routine clinical data, which may be noisier or incomplete, remains uncertain. Class imbalance in the dataset introduces challenges in threshold selection and may affect clinical applicability. Some predictive features may partly reflect definitional aspects of the weaning process. For instance, variables such as sedation state and oxygenation influence the timing of SA itself, potentially inflating their apparent predictive importance.

The dominant role of the timing of first SA, while informative, limits direct clinical actionability, as it reflects accumulated patient trajectory rather than an adjustable parameter. Additionally, exclusion criteria removed certain high-risk patient groups, such as those with early tracheostomy or unclear decision points, potentially limiting generalisability. Other limitations include the lack of analysis of temporal trends in variables such as fluid balance and uncertainty regarding applicability to tracheostomy-dependent populations. Despite these constraints, the spontaneous breathing trial remains the standard for assessing extubation readiness, and clinical judgement continues to guide decision-making.

Future Directions

Future work should prioritise external and prospective validation in diverse ICU populations. Integration of ML models into clinical workflows will require evaluation of their impact on clinician behaviour and patient outcomes. For example, model outputs could be used to prompt targeted interventions, such as sedation adjustment or optimisation of ventilator settings, prior to separation attempts. Prior randomised trials of computerised decision support for ventilator weaning have been either rule-based [23] or have evaluated closed-loop and model-driven systems in paediatric [24] or single-centre adult populations; an adult multicentre interventional trial of a contemporary ML tool embedded weaning workflow is a logical next step. Advances in high-frequency physiological monitoring may also enable more dynamic and continuous assessment of weaning readiness, complementing the static snapshot provided at the time of first SA.

Conclusions

Machine learning models demonstrated good performance in distinguishing easy from difficult or prolonged weaning trajectories, with more modest accuracy in predicting successful first extubation. Tabular foundation models achieved the strongest results. The most influential predictors were clinically interpretable and included timing of first SA, agitation burden, sedation state, organ dysfunction severity, frailty, and key respiratory parameters. These findings support the potential role of ML as an interpretable decision-support tool that highlights modifiable factors and informs pre-separation optimisation, while preserving the central role of clinical judgement in weaning decisions.

Abbreviations

AP, average precision. AUC, area under the ROC curve. CRF, case report form. EBM, Explainable Boosting Machine. EHR, electronic health record. ICU, intensive care unit. MLP, multi-layer perceptron. ML, machine learning. MV, mechanical ventilation. PEEP, positive end-expiratory pressure. P/F ratio, PaO_2/FiO_2 ratio. ROC, receiver operating characteristic. RSBI, rapid shallow breathing index. SA, separation attempt. SBT, spontaneous breathing trial. SHAP, SHapley Additive exPlanations. SOFA, Sequential Organ Failure Assessment. WEC, weaning eligibility criteria. WIND, WeanINg According to a New Definition.

Declarations

Ethics approval: This secondary analysis of the WEAN SAFE dataset was performed in accordance with the Declaration of Helsinki and applicable local and national regulations. The original WEAN SAFE study received ethics committee approval at participating ICUs, and consent was obtained where required or waived by the relevant ethics committee. The study was registered at ClinicalTrials.gov (NCT03255109).

Consent for publication: Not Applicable

Availability of data: The data are owned by contributing institutions. Requests should be made to the WEAN SAFE Executive Committee via the corresponding author.

Authors' contributions

Concept and design: NK, MM, JGL. Acquisition, analysis, or interpretation of data: All Authors. Drafting of the manuscript: RC, NK, BAM, MM, JGL. Critical revision: All Authors. Statistical analysis: NK, EU, MM. Supervision: MM, JGL.

Funding: Jointly funded by ESICM and the WEAN SAFE CRC from the European Respiratory Society.

Acknowledgments: The WEAN SAFE study was initiated by the Acute Respiratory Failure Section of ESICM. The authors thank Guy Francois from ESICM and the European Lung Foundation.

Competing Interests: NK, EU, RC, BAM, TP have no competing interests. LH reports grants from Liberate Medical and InflaRx, honoraria from Getinge. GB reports a grant from Drager, consulting fees from Flowmeter, honoraria from Drager and Getinge, stock options in Dico technologies. LB reports grants from Medtronic, Drager and Stimity, equipment from Fisher Paykel. MM reports fees from Fisher & Paykel and Lungpacer Medical. JGL reports grants from SFI and HRB Ireland, consulting fees from Cellenkos.

References

1. Pham T, Heunks L, Bellani G, Madotto F, Aragao I, Beduneau G, Goligher EC, Grasselli G, Laake JH, Mancebo J, et al. Weaning from mechanical ventilation in intensive care units across 50 countries (WEAN SAFE): a multicentre, prospective, observational cohort study. Lancet Respir Med. 2023;11(5):465–76.
2. Epstein SK, Ciubotaru RL. Independent effects of etiology of failure and time to reintubation on outcome for patients failing extubation. Am J Respir Crit Care Med. 1998;158(2):489–93.
3. Thille AW, Richard JC, Brochard L. The decision to extubate in the intensive care unit. Am J Respir Crit Care Med. 2013;187(12):1294–302.
4. Yang KL, Tobin MJ. A prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. N Engl J Med. 1991;324(21):1445–50.
5. Meade M, Guyatt G, Cook D, Griffith L, Sinuff T, Kergl C, Mancebo J, Esteban A, Epstein S. Predicting success in weaning from mechanical ventilation. Chest. 2001;120(6 Suppl):S400–24.
6. Karthika M, Al Enezi FA, Pillai LV, Arabi YM. Rapid shallow breathing index. Ann Thorac Med. 2016;11(3):167–76.
7. Roberts KJ, Goodfellow LT, Battey-Muse CM, Hoerr CA, Carreon ML, Sorg ME, Glogowski J, Girard TD, MacIntyre NR. Hess DR: AARC Clinical Practice Guideline: Spontaneous Breathing Trials for Liberation From Adult Mechanical Ventilation. Respir Care. 2024;69(7):891–901.
8. Godard S, Herry C, Westergaard P, Scales N, Brown SM, Burns K, Mehta S, Jacono FJ, Kubelik D, Maziak DE et al. Practice Variation in Spontaneous Breathing Trial Performance and Reporting. *Can Respir J* 2016, 2016:9848942.

9. Burns KEA, Khan J, Phoophiboon V, Trivedi V, Gomez-Builes JC, Giammarioli B, Lewis K, Chaudhuri D, Desai K, Friedrich JO. Spontaneous Breathing Trial Techniques for Extubating Adults and Children Who Are Critically Ill: A Systematic Review and Meta-Analysis. JAMA Netw Open. 2024;7(2):e2356794.
10. Beduneau G, Pham T, Schortgen F, Piquilloud L, Zogheib E, Jonas M, Grelon F, Runge I, Nicolas T, Grange S, et al. Epidemiology of Weaning Outcome according to a New Definition. The WIND Study. Am J Respir Crit Care Med. 2017;195(6):772–83.
11. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y. LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*: 2017; 2017.
12. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2016: 785–794.
13. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*: 2018; 2018.
14. Hollmann N, Muller S, Purucker L, Krishnakumar A, Korfer M, Hoo SB, Schirrmeister RT, Hutter F. Accurate predictions on small data with a tabular foundation model. Nature. 2025;637(8045):319–26.
15. Qu J, Holzmüller D, Varoquaux G, Le Morvan M. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In: *ICML 2025 - Forty-Second International Conference on Machine Learning*: 2025; 2025.
16. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2013: 623–631.
17. Jia Y, Kaul C, Lawton T, Murray-Smith R, Habli I. Prediction of weaning from mechanical ventilation using Convolutional Neural Networks. Artif Intell Med. 2021;117:102087.
18. Park JE, Kim DY, Park JW, Jung YJ, Lee KS, Park JH, Sheen SS, Park KJ, Sunwoo MH, Chung WY. Development of a Machine Learning Model for Predicting Weaning Outcomes Based Solely on Continuous Ventilator Parameters during Spontaneous Breathing Trials. Bioeng (Basel) 2023, 10(10).
19. Zappala S, Scaravilli V, Rovati L, Bosone M, Alfieri F, Ancona A, Grasselli G. Development and validation of a machine learning model for real-time prediction of invasive mechanical ventilation weaning readiness. J Crit Care. 2025;89:155105.
20. Laffey CM, Sheerin R, Khazaei O, McNicholas BA, Pham T, Heunks L, Bellani G, Brochard L, Tomescu D, Simpkin AJ, et al. Impact of frailty and older age on weaning from invasive ventilation: a secondary analysis of the WEAN SAFE study. Ann Intensive Care. 2025;15(1):13.
21. Igarashi Y, Ogawa K, Nishimura K, Osawa S, Ohwada H, Yokobori S. Machine learning for predicting successful extubation in patients receiving mechanical ventilation. Front Med (Lausanne). 2022;9:961252.

22. Zhao QY, Wang H, Luo JC, Luo MH, Liu LP, Yu SJ, Liu K, Zhang YJ, Sun P, Tu GW, et al. Development and Validation of a Machine-Learning Model for Prediction of Extubation Failure in Intensive Care Units. Front Med (Lausanne). 2021;8:676343.
23. Hsu JC, Chen YF, Chung WS, Tan TH, Chen T, Chiang JY. Clinical verification of a clinical decision support system for ventilator weaning. Biomed Eng Online. 2013;12(Suppl 1Suppl 1):S4.
24. Khemani RG, Bhalla A, Hotz JC, Klein MJ, Kwok J, Kohler K, Bornstein D, Chang D, Armenta-Quiroz A, Vu K, et al. Randomized Trial of Lung and Diaphragm Protective Ventilation in Children. NEJM Evid. 2025;4(6):EVIDoa2400360.

Table 1

Table 1 is available in the Supplementary Files section.