

Frailty as a predictor of outcome in non-invasive respiratory support for acute respiratory failure: a multicentre study

Lukasz Regulski

University of Warwick

Sina Saffaran

University of Warwick

Hang Yu

University of Warwick

Liam Weaver

University of Warwick

Marco Simonini

University of Modena Reggio-Emilia

Alessia Verduri

University of Modena Reggio-Emilia

Catherine Henry

University Hospitals of North Midlands NHS Trust

Abdisamad Ali

University Hospitals of North Midlands NHS Trust

Timothy E. Scott

University Hospitals of North Midlands NHS Trust

Naveed Mustfa

University Hospitals of North Midlands NHS Trust

Ajit Thomas

University Hospitals of North Midlands NHS Trust

Ashwin Rajhan

University Hospitals of North Midlands NHS Trust

Sannan Irshad

University Hospitals of North Midlands NHS Trust

Rehab Haider

University Hospitals of North Midlands NHS Trust

Jonathan Hewitt

Cardiff University

Keith Couper

University of Warwick

Enrico Clini

University of Modena Reggio-Emilia

Roberto Tonelli

rtonelli@unimore.it

University of Modena Reggio-Emilia

Declan G. Bates

University of Warwick

Research Article

Keywords: frailty, Clinical Frailty Scale, non-invasive respiratory support, high-flow nasal cannula, non-invasive ventilation, acute respiratory failure, machine learning, clinical prediction model, patient context

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10751751/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.

Failure of non-invasive respiratory support (NIRS) in acute respiratory failure is associated with poorer outcomes. Current clinical indices and proposed machine learning (ML) models that can be used to predict NIRS failure use measurements at initiation and during NIRS, but do not capture patients' a priori physiological reserves. Frailty, which is easily measured via the Clinical Frailty Scale (CFS), summarises a patient's accumulated health trajectory, but is rarely recorded during pre-ICU respiratory support, and its predictive value in the context of NIRS is unclear.

Methods.

We conducted a retrospective analysis of data on 341 patients treated with NIRS from centres in Italy and the UK. NIRS failure was defined as treatment escalation, death within seven days, or pre-specified physiological failure criteria. Frailty was assessed retrospectively by investigators blinded to NIRS outcome using pre-admission functional information. ML models using age, respiratory rate, and FiO_2 before and one-hour after starting NIRS were trained with and without including CFS, in two structurally different frameworks (TabPFN and XGBoost), alongside a logistic regression model using CFS alone. Performance was estimated by repeated stratified 10-fold cross-validation; discrimination was compared by the corrected resampled t-test and a paired bootstrap on pooled out-of-fold predictions.

Results.

Frailty was associated with NIRS failure (OR 1.19 per CFS point, 95% CI 1.05–1.34, $p = 0.005$); whereas age was not (OR 0.99, $p = 0.237$). CFS alone showed poor discrimination (AUC 0.584, 95% CI 0.523–0.645), but its inclusion improved discrimination in both ML models: TabPFN AUC increased from 0.77 to 0.80 ($\Delta + 0.04$, 95% CI + 0.01 to + 0.07, $p = 0.018$); XGBoost from 0.76 to 0.79 ($\Delta + 0.04$, 95% CI + 0.00 to + 0.07, $p = 0.043$). The improvement was positive in all 50 cross-validation partitions for both models. CFS was the second most influential feature in both models, after respiratory rate. Failure rates were flat across age quartiles while mean CFS rose from 2.8 to 5.1.

Conclusions.

Frailty may carry useful additional prognostic information for NIRS outcome when combined with other routinely collected measurements in predictive models. Although these hypothesis-generating findings require prospective validation, they support more widespread recording of the easily-measured CFS in acute respiratory care.

Trial registration.

Not applicable; retrospective analysis of existing clinical data.

Background

Non-invasive respiratory support (NIRS) — encompassing high-flow nasal cannula (HFNC) therapy, continuous positive airway pressure and non-invasive ventilation (NIV) — is the first-line treatment for many patients with acute respiratory failure (ARF). When it succeeds, it avoids or shortens intensive care unit (ICU) admission, avoids the risks of intubation, and is associated with lower mortality [1,2]. When it fails, however, patients whose treatment is escalated late appear to fare worse than those intubated earlier, plausibly because clinicians persist with a failing strategy for too long [2–4]. Identifying, in the first hours of treatment, those patients who are at higher risk of failing NIRS is therefore an important clinical problem.

The tools that are currently available for this purpose such as the ROX index [5], the HACOR score [6], and recently proposed machine learning (ML) models [7,8] share a common structure. They combine measurements of respiratory rate (RR), oxygenation and, in some cases, gas exchange, taken at or shortly after the initiation of NIRS. They are, in other words, descriptions of the patient's acute respiratory failure and its response to NIRS, rather than any representation of the underlying condition of the patient who is being treated.

This limitation matters because two patients can present with identical respiratory rates and oxygen requirements and yet have entirely different capacities to sustain the work of breathing required during NIRS over the following hours. Frailty is a clinical construct that can capture that difference: a state of reduced physiological reserve that increases vulnerability to adverse outcomes after a minor stressor [9]. It is measured most commonly with the Clinical Frailty Scale (CFS), a 9-point judgement-based scale grading functional dependency and overall health status, with scores of 5 to 8 indicating increasing frailty and 9 terminal illness [11]. The CFS is a compressed summary of the accumulated consequences of chronic disease, physical inactivity, nutrition, social circumstances and how consistently the patient has been able to look after their own health. It requires no equipment or blood sampling and can be assessed in a matter of minutes.

Frailty as measured by the CFS has been strongly associated with mortality and other adverse outcomes in critical care generally [10,24], and one previous study in elderly ICU patients found frailty to be associated with NIV application problems, NIV failure and mortality [12]. However, its incremental predictive value alongside early physiological responses for the specific task of identifying patients at risk of NIRS failure has not been established. We therefore aimed to determine whether frailty contributes predictive information beyond age and routinely measured early physiological variables in the early prediction of NIRS outcome. We addressed this with a feature-ablation design i.e. running otherwise identical models with and without CFS, in a pooled multicentre cohort [13–19], and we

deliberately repeated the entire analysis in two structurally unrelated machine learning model families so that any observed effect could be attributed to the variable rather than to the choice of algorithm.

Methods

Design and setting

This was a multicentre retrospective pooled analysis of prospectively collected data from seven clinical studies [13–19] conducted in the respiratory ICUs of the University Hospital of Modena Policlinico, Italy, and University Hospital of North Midlands NHS Trust, Stoke-on-Trent, UK. The study is reported with reference to the TRIPOD + AI reporting guideline for clinical prediction model studies [25].

Participants

The pooled dataset comprised 341 adults who received either NIV or HFNC therapy for ARF: 277 from University Hospital of Modena and 64 from University Hospital of North Midlands. The cohort was deliberately broad, including immunocompromised and immunocompetent patients with hypoxaemic ARF due to pneumonia, sepsis or COVID-19, and hypercapnic ARF due to acute exacerbation of chronic obstructive pulmonary disease or obesity hypoventilation syndrome. No patient contributed more than one episode.

Outcome

NIRS failure was defined, consistently across the contributing studies, as treatment escalation, death within seven days, or fulfilment of pre-specified physiological failure criteria even in the absence of escalation.

Frailty assessment

Frailty was assessed retrospectively by investigators blinded to NIRS outcome, according to published CFS criteria [11], using pre-admission functional information documented in ventilation unit admission proformas and in medical and nursing records. CFS grading reflected the most recent pre-morbid functional information available, not the patient's state during the acute episode. CFS was recorded for all 341 patients.

Physiological predictor variables

Physiological measurements were taken at two predefined time points: T0, within six hours before NIRS initiation, and T1, one to two hours after starting NIRS. The models used age, respiratory rate and fraction of inspired oxygen (FiO₂) at both time points, as well as the T1 – T0 change in RR and FiO₂.

Arterial blood gas variables (PaO_2 , PaCO_2) and the $\text{PaO}_2/\text{FiO}_2$ ratio were recorded and are reported in the cohort description, but were deliberately excluded from the models for two reasons. First, they were unavailable for the North Midlands cohort (64/341 patients, 18.8%), so including them would have forced either the exclusion of that cohort or the imputation of a fifth of the data on a variable known to be strongly outcome-associated. Second, the intended setting for this kind of prediction is pre-ICU, where a decision is often needed before an arterial sample is available. Restricting the feature set to eight variables also limits the risk of overfitting in a dataset of this size.

Missing data

All eight model features — CFS, age, respiratory rate and FiO_2 at both time points, and their T1 – T0 changes — were complete for all 341 patients, so no imputation was required and no imputation model could leak information between folds. The arterial blood gas variables were unavailable for the 64 North Midlands patients and are reported for cohort description only; they were not used as model features. The CFS-only logistic regression had no missing values to handle.

Models and feature-ablation design

Three feature configurations were evaluated (1) CFS alone, in a logistic regression model, to establish whether frailty is usable as a standalone predictor; (2) Age and early physiology measurements without CFS; (3) Age and early physiology measurements with CFS. Configurations 2 and 3 were each run in two independent ML frameworks. The primary model was TabPFN, a transformer-based foundation model for tabular data that performs inference in a single forward pass and has shown strong performance on small clinical datasets without hyperparameter tuning [20], including in NIRS outcome prediction [18,19]. The second was XGBoost, a gradient-boosting framework that builds an ensemble of decision trees sequentially [22]. These two model classes make very different assumptions; agreement between them is therefore informative about whether an effect is due to the underlying data or to the algorithm. No hyperparameter tuning was performed for either model; both were used with fixed default-like settings (TabPFN: 32 estimators; XGBoost: 100 trees, learning rate 0.1, maximum depth 4).

Validation and performance metrics

Performance was estimated using repeated stratified 10-fold cross-validation. The complete procedure was repeated over 50 independent fold partitions generated from 50 different random seeds, so that no result depends on one particular split of the data; within each partition every model was evaluated on exactly the same folds, so that all model comparisons are paired. Models were fitted on 90% of the data and evaluated on the held-out 10%; no model was evaluated on data it had been fitted on. Discrimination was quantified by the area under the receiver operating characteristic curve (AUC), which is independent of any classification threshold, and overall probabilistic prediction accuracy by the Brier score, which combines calibration and discrimination in a single value. Accuracy, balanced accuracy, sensitivity and

specificity were additionally reported at a Youden-optimal classification threshold. To avoid selecting an operating point on the data used to score it, the threshold applied to each held-out fold was the value maximising the Youden index on the pooled out-of-fold predictions of the remaining nine folds. Equivalent results at a fixed 0.5 threshold are given in Additional file 1.

Statistical analysis

Continuous baseline variables are reported as mean $\pm$ standard deviation and were compared between outcome groups with independent two-sample Student's t-tests; categorical variables were compared with Pearson's chi-square test. Unadjusted associations between individual variables and outcome are reported as odds ratios with 95% confidence intervals from univariable logistic regression.

Models were compared in two complementary ways. The primary comparison was the corrected resampled t-test of Nadeau and Bengio applied to the 500 per-fold differences in AUC, which inflates the variance estimate to account for the overlap between training sets that an ordinary paired t-test ignores. Second, the out-of-fold predicted probabilities were pooled within each partition to give one held-out prediction per patient, and the difference in AUC was tested with a paired stratified bootstrap in which the same resampled patients were used for both models; resamples were drawn within each partition and pooled across all 50, so that the resulting interval carries both patient-sampling and partition-to-partition variability. Confidence intervals for individual AUCs were obtained in the same way. No adjustment for multiple comparisons was applied; the two model families address the same pre-specified question and are reported jointly rather than as independent tests.

Model interpretation

Feature contributions were quantified with SHAP (SHapley Additive exPlanations) values [21], computed on held-out data within each fold and aggregated across folds. For TabPFN, Kernel SHAP was used with 30 background samples selected by k-means clustering and 1000 perturbation samples per test instance; for XGBoost, Tree SHAP was used, which yields exact Shapley values for tree ensembles. A positive SHAP value indicates that the feature pushed that patient's prediction towards NIRS failure relative to the cohort average.

Results

Cohort characteristics

The pooled cohort comprised 341 patients (192 HFNC, 149 NIV), of whom 137 (40.2%) met the definition of NIRS failure. Mean age was 67.6 ± 12.7 years and mean CFS was 3.8 ± 1.9 . Baseline characteristics by outcome group are shown in Table 1. Respiratory rate and FiO_2 at both time points, and the $\text{PaO}_2/\text{FiO}_2$ ratio, differed significantly between patients who succeeded and failed NIRS; PaO_2 and PaCO_2 did not.

Table 1

Baseline demographics and physiological variables at T0 and T1, overall, by respiratory support modality, and by NIRS outcome.

Variable	Overall (N = 341)	HFNC (n = 192)	NIV (n = 149)	Success (n = 204)	Failure (n = 137)	pt
Age, years	67.6 ± 12.7	68.0 ± 13.1	67.2 ± 12.3	68.3 ± 12.1	66.6 ± 13.6	0.237
Male sex, n (%)	188 (67.9)	127 (66.1)	61 (71.8)	101 (64.3)	87 (72.5)	0.189
Clinical Frailty Scale	3.8 ± 1.9	3.3 ± 1.8	4.4 ± 1.7	3.5 ± 2.0	4.1 ± 1.6	0.005
CFS ≥ 5, n (%)	128 (37.5)	51 (26.6)	77 (51.7)	68 (33.3)	60 (43.8)	0.065
NIRS failure, n (%)	137 (40.2)	66 (34.4)	71 (47.7)	—	—	
Respiratory rate, breaths/min (T0)	27.8 ± 7.6	27.0 ± 5.7	28.8 ± 9.4	26.1 ± 6.6	30.4 ± 8.2	< 0.001
PaO ₂ , mmHg (T0)	66.4 ± 10.4	65.6 ± 10.5	68.1 ± 10.0	66.6 ± 10.0	66.1 ± 11.0	0.678
FiO ₂ , % (T0)	51.3 ± 15.9	51.6 ± 13.7	50.9 ± 18.4	48.2 ± 16.1	55.8 ± 14.5	< 0.001
PaO ₂ /FiO ₂ ratio, mmHg (T0)	134.5 ± 42.5	138.7 ± 44.0	125.0 ± 37.4	142.6 ± 45.5	123.9 ± 35.8	< 0.001
PaCO ₂ , mmHg (T0)	35.9 ± 9.2	36.8 ± 9.8	33.8 ± 7.1	35.9 ± 9.5	35.9 ± 8.8	0.987
mROX index (T0)	5.0 ± 2.2	5.4 ± 2.2	4.1 ± 1.7	5.4 ± 2.2	4.4 ± 2.0	< 0.001
Respiratory rate, breaths/min (T1)	24.4 ± 6.2	23.7 ± 5.0	25.3 ± 7.4	22.6 ± 5.4	27.2 ± 6.4	< 0.001
PaO ₂ , mmHg (T1)	66.5 ± 8.2	66.2 ± 8.9	67.0 ± 6.2	65.9 ± 8.3	67.3 ± 8.0	0.155
FiO ₂ , % (T1)	50.1 ± 15.4	50.3 ± 15.7	50.0 ± 15.0	45.9 ± 14.1	56.5 ± 15.2	< 0.001

Values are mean ± standard deviation unless otherwise stated.

† Success versus failure; independent two-sample Student's t-test for continuous variables, Pearson's chi-square test for categorical variables.

‡ Sex was not recorded for the 64 patients of the North Midlands cohort, all of whom received NIV; percentages are of the 277 patients with sex recorded.

§ Arterial blood gas variables and the derived PaO₂/FiO₂ ratio were unavailable for the 64 patients of the North Midlands cohort and are reported for descriptive purposes only; they were not used as model features (see Methods).

Variable	Overall (N = 341)	HFNC (n = 192)	NIV (n = 149)	Success (n = 204)	Failure (n = 137)	p†
PaO ₂ /FiO ₂ ratio, mmHg (T1)	142.4 ± 40.9	144.9 ± 43.4	136.7 ± 34.0	155.5 ± 41.6	125.3 ± 32.9	< 0.001
PaCO ₂ , mmHg (T1)	36.8 ± 8.6	37.9 ± 9.5	34.5 ± 5.7	37.3 ± 7.5	36.2 ± 9.9	0.320
mROX index (T1)	6.1 ± 2.7	6.4 ± 2.7	5.2 ± 2.5	6.9 ± 2.6	5.0 ± 2.4	< 0.001
Values are mean ± standard deviation unless otherwise stated.						
† Success versus failure; independent two-sample Student's t-test for continuous variables, Pearson's chi-square test for categorical variables.						
‡ Sex was not recorded for the 64 patients of the North Midlands cohort, all of whom received NIV; percentages are of the 277 patients with sex recorded.						
§ Arterial blood gas variables and the derived PaO ₂ /FiO ₂ ratio were unavailable for the 64 patients of the North Midlands cohort and are reported for descriptive purposes only; they were not used as model features (see Methods).						

Frailty and age as standalone predictors

Frailty was associated with NIRS failure (odds ratio 1.19 per CFS point, 95% CI 1.05–1.34, $p = 0.005$; mean CFS 4.1 ± 1.6 in failures versus 3.5 ± 2.0 in successes). Age was not (odds ratio 0.99 per year, 95% CI 0.97–1.01, $p = 0.237$). Frailty defined dichotomously as CFS ≥ 5 was present in 43.8% of failures and 33.3% of successes ($p = 0.065$). Failure rates were flat across age quartiles while mean CFS rose from 2.8 to 5.1 (Table S1). A logistic regression model using CFS alone achieved a pooled out-of-fold AUC of 0.584 (95% CI 0.523–0.645). At its Youden-optimal operating point it reached a sensitivity of 0.725 but at the cost of a specificity of 0.382; no threshold yielded a balanced combination of the two, because the underlying discrimination is poor. Frailty as a static measure of baseline vulnerability is not, on its own, a reliable predictor of the dynamic physiological trajectory that determines NIRS outcome.

Inclusion of frailty in machine learning models adds non-redundant information

Adding CFS to age and early physiology improved discrimination in both machine learning model families (Table 2, Fig. 1). For TabPFN, pooled out-of-fold AUC rose from 0.767 (95% CI 0.713–0.819) to 0.804 (0.754–0.852), a paired difference of +0.037 (95% CI +0.008 to +0.067, $p = 0.018$). For XGBoost, AUC rose from 0.759 (0.703–0.813) to 0.794 (0.742–0.844), a difference of +0.035 (95% CI +0.001 to +0.070, $p = 0.043$). The improvement was in the same direction in every one of the 50 cross-validation partitions for both frameworks, and in 393 of 500 individual folds for TabPFN and 379 of 500 for XGBoost. Inclusion of CFS also improved overall probabilistic prediction accuracy in both frameworks, with the Brier score falling from 0.189 ± 0.031 to 0.174 ± 0.033 for TabPFN ($p = 0.011$) and from 0.204 ± 0.040 to 0.185 ± 0.041 for XGBoost ($p = 0.026$); calibration plots are shown in Additional file 1, Figure S1.

Table 2. Predictive performance for the CFS-only logistic regression and for each machine learning model with and without CFS. Repeated stratified 10-fold cross-validation (N=341, 50 partitions); identical folds for all models within each partition.

Panel A – TabPFN

Metric	CFS alone	Without CFS	With CFS	p, with vs without CFS
AUC (95% CI)	0.584 (0.523–0.645)	0.767 (0.713–0.819)	0.804 (0.754–0.852)	0.018†‡
Accuracy	0.519 ± 0.068	0.719 ± 0.073	0.736 ± 0.068	0.401
Balanced accuracy	0.553 ± 0.068	0.711 ± 0.073	0.740 ± 0.068	0.152
Sensitivity	0.725 ± 0.210	0.667 ± 0.132	0.758 ± 0.126	0.033
Specificity	0.382 ± 0.176	0.754 ± 0.114	0.722 ± 0.108	0.385
Brier score	0.237 ± 0.012	0.189 ± 0.031	0.174 ± 0.033	0.011

Panel B – XGBoost

Metric	CFS alone	Without CFS	With CFS	p, with vs without CFS
AUC (95% CI)	0.584 (0.523–0.645)	0.759 (0.703–0.813)	0.794 (0.742–0.844)	0.043†‡
Accuracy	0.519 ± 0.068	0.698 ± 0.071	0.737 ± 0.072	0.080
Balanced accuracy	0.553 ± 0.068	0.708 ± 0.071	0.740 ± 0.072	0.137
Sensitivity	0.725 ± 0.210	0.754 ± 0.129	0.754 ± 0.126	0.993
Specificity	0.382 ± 0.176	0.661 ± 0.113	0.726 ± 0.114	0.094
Brier score	0.237 ± 0.012	0.204 ± 0.040	0.185 ± 0.041	0.026

Values are mean ± standard deviation across all 500 folds (50 partitions × 10 folds) unless otherwise stated. Accuracy, balanced accuracy, sensitivity and specificity are computed at a Youden-optimal threshold selected on the pooled out-of-fold predictions of the other nine folds; AUC and Brier score are threshold-independent. Lower Brier scores indicate better overall probabilistic prediction accuracy.

† Corrected resampled t-test (Nadeau–Bengio) across the 500 folds.

‡ Difference in pooled out-of-fold AUC from adding CFS: +0.037 (95% CI +0.008 to +0.067) for TabPFN and +0.035 (95% CI +0.001 to +0.070) for XGBoost, from a paired stratified bootstrap drawn within each

partition and pooled across all 50 (bootstrap $p=0.014$ and $p=0.047$ respectively).

The threshold-dependent metrics moved in the same direction but were less clear-cut. At the Youden-optimal cut-off, balanced accuracy improved from 0.711 ± 0.073 to 0.740 ± 0.068 for TabPFN and from 0.708 ± 0.071 to 0.740 ± 0.072 for XGBoost, but neither difference reached statistical significance ($p = 0.152$ and $p = 0.137$). For TabPFN the gain was concentrated in sensitivity (0.667 to 0.758, $p = 0.033$) and for XGBoost in specificity (0.661 to 0.726, $p = 0.094$).

SHAP analysis of the TabPFN model including CFS (Fig. 2A) ranked respiratory rate at T1 first and CFS second, with higher CFS scores pushing predictions towards failure. The same ordering was found for XGBoost (Fig. 3A).

Discussion

The principal finding of this study is the consistent observation, across two structurally unrelated machine learning frameworks applied to a diverse multicentre cohort, that frailty contributes non-redundant prognostic information beyond age and early physiological response in the early prediction of NIRS outcome.

Our results are consistent with the broader literature on ageing and frailty as distinct constructs [23] and with the argument that frailty in critical care must be measured rather than inferred [10,24]. Age and CFS were only moderately correlated (Pearson $r = 0.45$), and NIRS failure rates were essentially flat across age quartiles (40.4%, 41.6%, 39.0%, 39.7% from youngest to oldest) while mean CFS rose monotonically from 2.8 to 5.1 across the same quartiles.

Interestingly, although age contributed little predictive information to the machine learning models when frailty was absent, it became the third most important feature once frailty was included, with older age associated with predictions of NIRS success. We can offer two non-exclusive potential explanations for this counter-intuitive result. The first is aetiological: older patients in this cohort are more likely to present with acute-on-chronic conditions such as COPD exacerbation or cardiogenic pulmonary oedema, which respond comparatively well to NIRS, whereas younger patients are more likely to present with severe acute hypoxaemic events such as pneumonia or ARDS, which do not. Age, in a model that already contains frailty, may thus be acting as a proxy for the underlying aetiology. The second relates to treatment allocation: younger patients without limitations of care are more likely to be escalated to invasive support, and escalation is part of our composite failure definition, so some of the age effect may reflect clinician behaviour rather than patient physiology.

Although the size of the effect of including frailty on predictive accuracy is modest in absolute terms (roughly 0.04 of AUC), it is notable that adding one easily and cheaply measured variable to models from two entirely different machine learning model families (a) increased model discrimination performance to the threshold often associated with clinical utility (AUC = 0.8) [26], (b) improved the balance between the sensitivity and specificity of model predictions, and (c) improved overall probabilistic prediction

accuracy (lower Brier score). Our findings are also consistent with the single previous study of frailty and NIV outcome in elderly ICU patients, which reported associations between CFS and NIV failure and mortality [12], and with meta-analytic evidence that frailty predicts mortality and adverse outcomes across critical care populations [24].

A methodological strength of this analysis is that the entire cross-validation procedure was repeated over 50 independent fold partitions rather than reported from a single split, and that model comparison used the corrected resampled t-test, which accounts for the overlap between cross-validation training sets. Estimates from a single partition of a dataset of this size can vary appreciably; reporting the distribution across partitions, and an interval that carries that variability, gives a more realistic picture of what the added value of frailty would be in a new sample.

Several limitations qualify these findings and warrant emphasis. All data were retrospective, and CFS in particular was assigned retrospectively from records rather than prospectively at the bedside. Retrospective CFS assignment is known to be less reliable than contemporaneous assessment, and although assessors were blinded to outcome, some misclassification is possible. This would in general bias towards underestimating frailty's contribution. The pooled dataset resulted in a heterogeneous cohort of patients treated with different forms of NIRS, and the resulting composite failure outcome included escalation, death and physiological criteria, thus capturing treatment decisions and ceiling of care as well as pure NIRS failure. Finally, because of the limited size of the dataset, all data were required for model development and no separate external validation cohort was held out. Performance estimates therefore come from cross-validation within a single pooled dataset, and generalisability remains unproven. These findings should therefore be regarded as hypothesis-generating.

Conclusions

In a pooled multicentre cohort of 341 patients receiving non-invasive respiratory support for acute respiratory failure, frailty measured with the Clinical Frailty Scale carried prognostic information that neither age nor early physiological response supplied. Although frailty was unreliable as a stand-alone predictor, including it in multivariable prediction models improved measures of predictive performance consistently across two structurally different machine learning frameworks and across 50 independent cross-validation partitions. More widespread assessment of frailty in the early stages of acute respiratory care could therefore make an important and easily achievable contribution to the development of more accurate decision-support tools to help clinicians identify patients at risk of failing NIRS.

Abbreviations

ARDS

acute respiratory distress syndrome

ARF

acute respiratory failure
AUC
area under the receiver operating characteristic curve
CFS
Clinical Frailty Scale
CI
confidence interval
COPD
chronic obstructive pulmonary disease
CPAP
continuous positive airway pressure
 FiO_2
fraction of inspired oxygen
HACOR
heart rate, acidosis, consciousness, oxygenation and respiratory rate score
HFNC
high-flow nasal cannula
ICU
intensive care unit
ML
machine learning
NIRS
non-invasive respiratory support
NIV
non-invasive ventilation
OR
odds ratio
 PaCO_2
partial pressure of arterial carbon dioxide
 PaO_2
partial pressure of arterial oxygen
 $\text{PaO}_2/\text{FiO}_2$
ratio of arterial oxygen partial pressure to fraction of inspired oxygen
ROX
ratio of oxygen saturation index
RR
respiratory rate
SD
standard deviation
SHAP

SHapley Additive exPlanations
T0
within six hours before NIRS initiation
T1
one to two hours after NIRS initiation
TabPFN
Tabular Prior-data Fitted Network
XGBoost
extreme gradient boosting.

Declarations

Ethics approval and consent to participate

This retrospective pooled analysis used previously collected clinical data from participating centres in Italy and the UK. The contributing studies conducted at the University Hospital of Modena were approved by the Comitato Etico Area Vasta Emilia Nord (Ethics Committee Area Vasta Emilia Nord), including registered protocol 4485/C.E. (document 266/16) and, for the later prospective cohort, protocol 165/2024, as applicable. Informed consent for participation and for the use of clinical data was obtained in the original Italian studies from participants or their relatives, as appropriate.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the UKRI Engineering and Physical Sciences Research Council (Ref. EP/W000490/1 and EP/Y003527/1) and The Royal Academy of Engineering (Ref. RF2122-21–258).

Author Contribution

All authors contributed to study conceptualization. LR performed the formal analysis, developed the methodology, implemented the software, and contributed to writing the manuscript. HY and SS contributed to methodology development and writing – review & editing. MS, AV, RT, EC, CH, AA, TES, NM,

AT, AR, SI, and RH generated patient datasets used in the study. DGB provided supervision and contributed to writing – original draft. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

Patient data used for internal validation and machine learning models are available upon request to bona fide researchers for specific scientific purposes. The analysis code, model configurations, random seeds, aggregate (non-identifiable) result tables and the scripts that generate every figure in this article are publicly available at <https://github.com/LukaszRegulski03/frailty-nirs-outcome>.

References

1. Grieco DL, Maggiore SM, Roca O, Spinelli E, Patel BK, Thille AW, et al. Non-invasive ventilatory support and high-flow nasal oxygen as first-line treatment of acute hypoxemic respiratory failure and ARDS. *Intensive Care Med*. 2021;47:851–66. doi:10.1007/s00134-021-06459-2.
2. Ferreyro BL, Angriman F, Munshi L, Del Sorbo L, Ferguson ND, Rochweg B, et al. Association of noninvasive oxygenation strategies with all-cause mortality in adults with acute hypoxemic respiratory failure. *JAMA*. 2020;324:57. doi:10.1001/jama.2020.9524.
3. Bellani G, Laffey JG, Pham T, Madotto F, Fan E, Brochard L, et al. Noninvasive ventilation of patients with acute respiratory distress syndrome. Insights from the LUNG SAFE study. *Am J Respir Crit Care Med*. 2017;195:67–77. doi:10.1164/rccm.201606-1306OC.
4. Battaglini D, Robba C, Ball L, Silva PL, Cruz FF, Pelosi P, et al. Noninvasive respiratory support and patient self-inflicted lung injury in COVID-19: a narrative review. *Br J Anaesth*. 2021;127:353–64. doi:10.1016/j.bja.2021.05.024.
5. Roca O, Caralt B, Messika J, Samper M, Sztrymf B, Hernández 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. doi:10.1164/rccm.201803-0589OC.
6. Duan J, Chen L, Liu X, Bozbay S, Liu Y, Wang K, et al. An updated HACOR score for predicting the failure of noninvasive ventilation: a multicenter prospective observational study. *Crit Care*. 2022;26:196. doi:10.1186/s13054-022-04060-7.
7. Yu H, Saffaran S, Tonelli R, Laffey JG, Esquinas AM, de Lima LM, et al. Machine learning models compared with current clinical indices to predict the outcome of high flow nasal cannula therapy in acute hypoxemic respiratory failure. *Crit Care*. 2025;29:101. doi:10.1186/s13054-025-05336-4.

8. Yu H, Saffaran S, Tonelli R, Laffey JG, Zhang Q, Esquinas AM, et al. Evaluating the effect of heart and respiratory rate measurement errors on the ability to predict the outcome of high flow nasal cannula therapy: a multi-centre study. *Crit Care*. 2025;29:535.
9. Clegg A, Young J. The frailty syndrome. *Clin Med*. 2011;11:72–5.
10. Ferrante LE, Szczeklik W. Frailty is crucial in FORECASTing outcomes in critical care. *Intensive Care Med*. 2024;50:1119–22. doi:10.1007/s00134-024-07518-0.
11. Rockwood K, Song X, MacKnight C, Bergman H, Hogan DB, McDowell I, et al. A global clinical measure of fitness and frailty in elderly people. *CMAJ*. 2005;173:489–95.
12. Kara I, Yildirim F, Zerman A, Gullu Z, Boyaci N, Aydogdu M, et al. The impact of frailty on noninvasive mechanical ventilation in elderly medical intensive care unit patients. *Aging Clin Exp Res*. 2018;30:359–66. doi:10.1007/s40520-017-0774-z.
13. Tonelli R, Fantini R, Tabbì L, Castaniere I, Pisani L, Pellegrino MR, et al. Early inspiratory effort assessment by esophageal manometry predicts noninvasive ventilation outcome in de novo respiratory failure. A pilot study. *Am J Respir Crit Care Med*. 2020;202:558–67. doi:10.1164/rccm.201912-2512OC.
14. Tonelli R, Busani S, Tabbì L, Fantini R, Castaniere I, Biagioni E, et al. Inspiratory effort and lung mechanics in spontaneously breathing patients with acute respiratory failure due to COVID-19: a matched control study. *Am J Respir Crit Care Med*. 2021;204:725–8. doi:10.1164/rccm.202104-1029LE.
15. Tonelli R, Cortegiani A, Fantini R, Tabbì L, Castaniere I, Biagioni E, 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;207:787–9. doi:10.1164/rccm.202210-1848LE.
16. Tonelli R, Bruzzi G, Fantini R, Tabbì L, Castaniere I, Andrisani D, et al. Assessment of nasal pressure swing predicts respiratory support dependency in patients with hypoxic respiratory failure. *Eur J Intern Med*. 2025;135:91–7. doi:10.1016/j.ejim.2025.02.016.
17. Tabbì L, Marchioni A, Messina J, Tonelli R, Castaniere I, Andrisani D, et al. Ultrasound-based evaluation of inspiratory muscle effort in patients with type 2 respiratory failure secondary to exacerbation of chronic obstructive pulmonary disease. *Eur J Intern Med*. 2025. doi:10.1016/j.ejim.2025.106649.
18. Yu H, Saffaran S, Maia IS, Clini E, Bates DG, NIVPredict Study Group. Early prediction of non-invasive ventilation outcome using the TabPFN machine learning model: a multi-centre validation study. *Intensive Care Med*. 2025;51:1542–4.
19. Yu H, Saffaran S, Ali A, Henry C, Scott TE, Mustafa N, et al. In-hospital testing of NIVPredict – an AI tool for early prediction of non-invasive ventilation outcome in acute respiratory failure. *Crit Care*. 2026;30:124. doi:10.1186/s13054-026-05894-1.
20. Hollmann N, Müller S, Purucker L, Krishnakumar A, Körfer M, Hoo SB, et al. Accurate predictions on small data with a tabular foundation model. *Nature*. 2025;637:319–26. doi:10.1038/s41586-024-08328-6.

21. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. *Clin Transl Sci*. 2024;17:e70056. doi:10.1111/cts.70056.
22. Moore A, Bell M. XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: a UK Biobank cohort study. *Clin Med Insights Cardiol*. 2022;16:11795468221133611. doi:10.1177/11795468221133611.
23. Gordon AL, Witham MD, Henderson EJ, Harwood RH, Masud T. Research into ageing and frailty. *Future Healthc J*. 2021;8:e237–42. doi:10.7861/fhj.2021-0088.
24. Muscedere J, Waters B, Varambally A, Bagshaw SM, Boyd JG, Maslove D, et al. The impact of frailty on intensive care unit outcomes: a systematic review and meta-analysis. *Intensive Care Med*. 2017;43:1105–22. doi:10.1007/s00134-017-4867-0.
25. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
26. Jeffreys E, Jenkinson A, Dassios T, Greenough A. Can artificial intelligence predict failure of non-invasive respiratory support in the neonatal unit?. *Semin Fetal Neonatal Med*. 2025. doi:10.1016/j.siny.2025.101683.

Figures

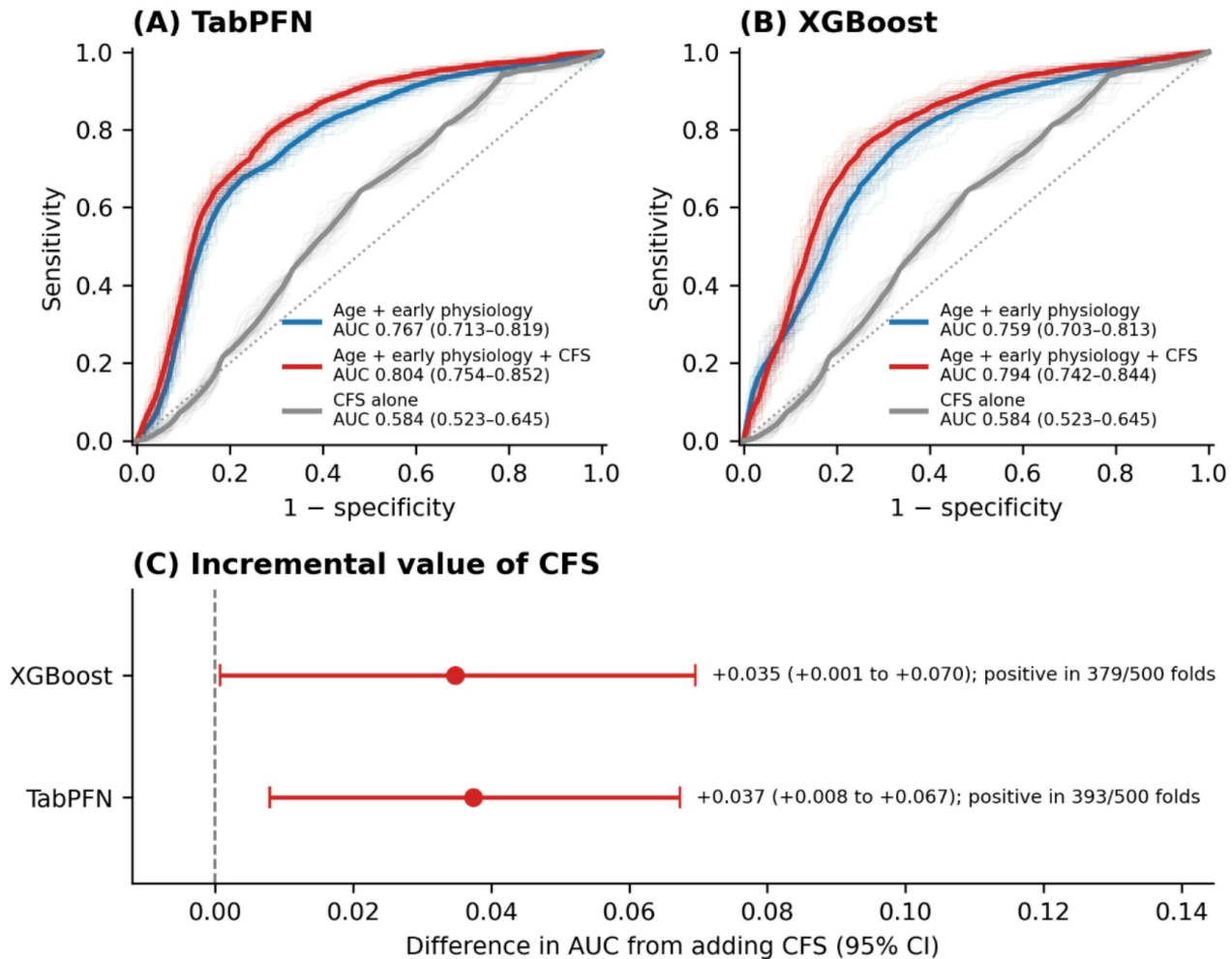
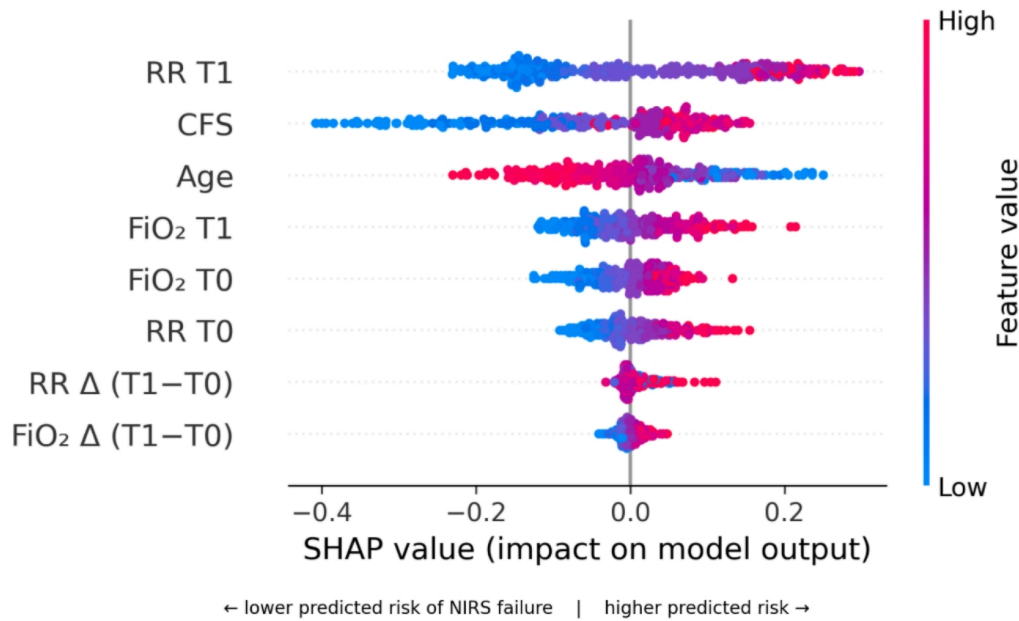


Figure 1

Discrimination for the prediction of non-invasive respiratory support failure (N=341, pooled out-of-fold predictions from repeated stratified 10-fold cross-validation, 50 partitions). (A) Receiver operating characteristic curves for the TabPFN model using CFS alone, age plus early physiology, and age plus early physiology plus CFS; faint lines are the 50 individual partitions and bold lines their mean. (B) The same comparison for XGBoost. (C) Difference in out-of-fold AUC attributable to adding CFS, with 95% confidence intervals from a paired stratified bootstrap drawn within each partition and pooled across all 50; the dashed line marks no difference.

(A) TabPFN including CFS



(B) TabPFN excluding CFS

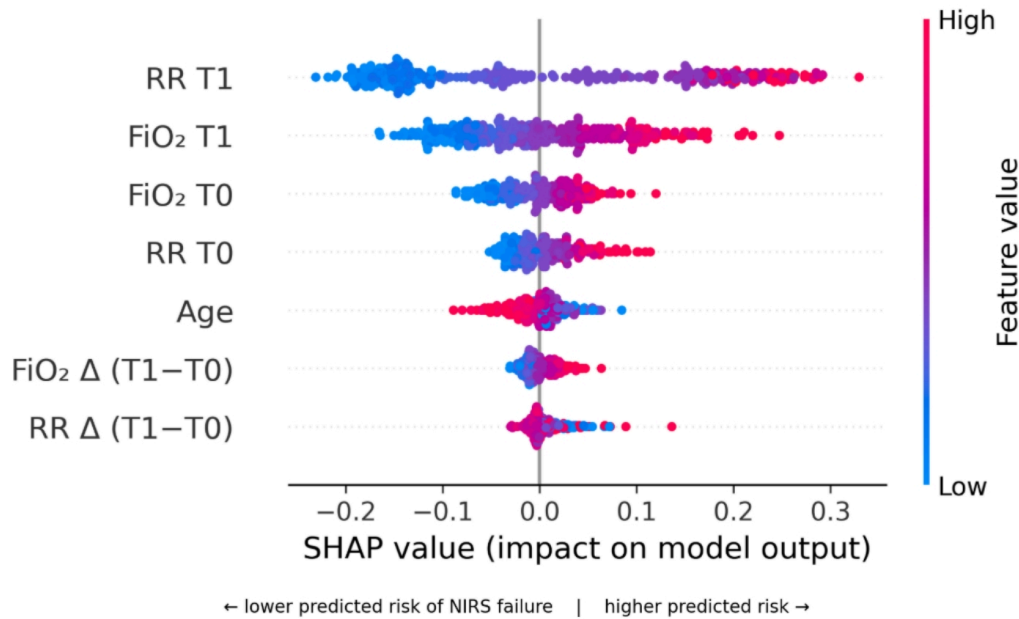
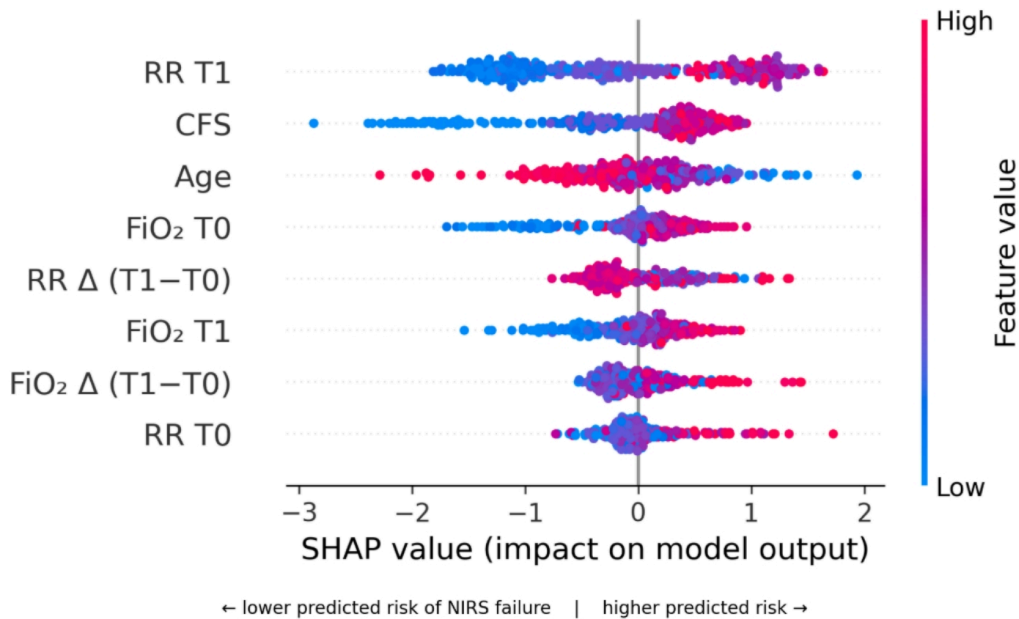


Figure 2

SHAP summary plots for the TabPFN model, computed on held-out data within each cross-validation fold and aggregated across folds. (A) Model including CFS. (B) Model excluding CFS. Features are ordered by mean absolute SHAP value. Each point is one patient; colour encodes that patient's value of the feature (blue low, red high). Positive SHAP values indicate that the feature increased the predicted probability of NIRS failure for that patient.

(A) XGBoost including CFS



(B) XGBoost excluding CFS

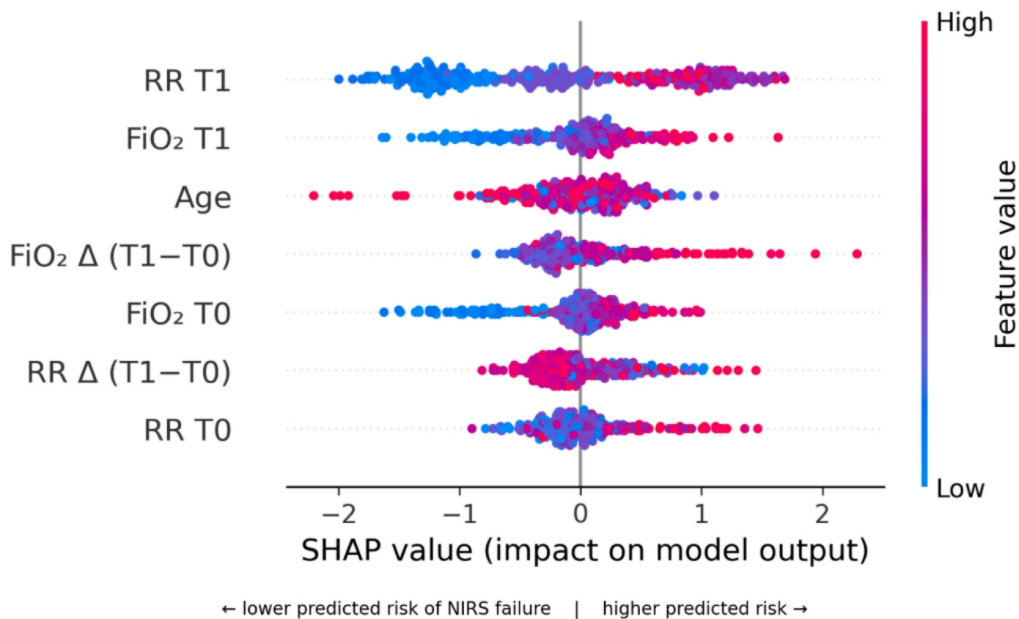


Figure 3

SHAP summary plots for the XGBoost model, computed with Tree SHAP on held-out data within each cross-validation fold and aggregated across folds. (A) Model including CFS. (B) Model excluding CFS. Plot conventions as in Figure 2.

Supplementary Files

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

- [Supplementarymaterials.docx](#)