Università degli Studi di Modena e Reggio Emilia

Facoltà di Medicina e Chirurgia

Corso di Laurea Magistrale in Medicina e Chirurgia

Dipartimento di Scienze Mediche e Chirurgiche, Materno-Infantili e dell'Adulto

Struttura Complessa di Malattie dell'Apparato Respiratorio

Direttore: Prof. Enrico Clini

Frailty as a predictor for outcome of Non-Invasive Respiratory Support in Acute Respiratory Failure: an innovative study supported by Machine Learning algorithms

RELATORE:

Prof. Roberto Tonelli

TESI DI LAUREA DI:

Marco Simonini

Anno Accademico 2025-2026

A mio zio Paolo,

alla pecorella sul tetto del presepe

Vivere, senza malinconia
Ridere, delle follie del mondo

Abstract

Introduction

Non-Invasive Respiratory Support (NIRS) is a crucial treatment option for Acute Respiratory Failure (ARF), though NIRS failure may lead to delayed advanced invasive support and therefore increased mortality. Frailty is a multidimensional syndrome due to depletion in physiologic reserves resulting in increased susceptibility to adverse outcomes and has been studied in the past in its relationship with ARF and NIRS outcome. Machine Learning (ML) is a field of Artificial Intelligence that provides output, such as failure probability, given a certain amount of input data.

Materials and Methods

This study was a multicentre retrospective analysis of data on 341 ARF patients treated with NIRS in Italy and in the UK. Primary aim was to investigate the predictive value of frailty in addition to conventional clinical data through the use of ML, exploratory analysis was to study its performance in specific subgroup, particularly the HFNC-only one. NIRS failure was defined as treatment escalation, death within seven days, or physiological failure. Frailty was assessed using the Clinical Frailty Scale (CFS) retrospectively using pre-admission functional information. Two ML models (TabPFN and XGBoost) were used, providing age and clinical data (respiratory rate and FiO_2) within 6 hours before NIRS initiation and 1-2 hours after, and through feature ablation the impact of adding pre-admission CFS was quantified. A logistic regression model using CFS alone was also used. Performance was quantified through repeated 10-fold cross-validation; Area Under the Curve (AUC), Sensitivity, Specificity, Accuracy, Balanced Accuracy and Brier Score were calculated and compared by the corrected t-test. SHapley Additive exPlanations (SHAP) analysis was used to show the weighted contribution of each input variable to the output.

Results

Frailty was highly prevalent in NIRS failure without statistical significance (AUC=0.59±0.07, p=0.065), but feature ablation proved that CFS was significantly influent in ML predictive performance increasing AUC from 0.77±0.08 to 0.82±0.07 in TabPFN (p=0.001) and from 0.77 ± 0.08 to 0.81 ± 0.06 for XGBoost (p=0.001).

SHAP analysis showed CFS to be the second most important determinant in both ML models. Exploratory analysis in HFNC-only subgroup showed an even more enhanced predictive performance through CFS inclusion (AUC increased from 0.889 ± 0.08 to 0.946 ± 0.035, p=0.009).

Conclusions

CFS appears to yield relevant influence on ML predictive performance, showing complementary properties to age and conventional clinical data. Combination of all these data may cover all stages of development of ARF, from baseline vulnerability to dynamic physiological response. Furtherly enhanced predictive performance in HFNC-only subgroup may support the hypothesis of interpreting HFNC as a NIRS form that is more dependent on residual functional capacity which is lower in frail patients. Although these hypothesis-generating findings require prospective validation, they support more widespread recording of CFS in acute respiratory care.

Summary

1 Introduction 1
1.1 Acute respiratory failure 1
1.1.1 Definition of Acute Respiratory Failure 1
1.1.2 Pathophysiological classification of Acute Respiratory Failure 1
1.1.2.1 Type 1 ARF 1
1.1.2.2 Type 2 ARF 3
1.1.2.3 Oxygen-induced hypercapnia 5
1.1.3 Clinical presentation of Acute Respiratory Failure 5
1.1.3.1 Clinical presentation of hypoxia in type 1 ARF 5
1.1.3.2 Clinical presentation of hypercapnia in type 2 ARF 6
1.1.4 Diagnosis of Acute Respiratory Failure 6
1.1.4.1 Pulse oximetry 6
1.1.4.2 Arterial Blood Gas analysis 6
1.1.4.3 Laboratory exams 7
1.1.4.4 Chest radiography 7
1.1.4.5 Ultrasound 7
1.1.4.6 Chest CT 8
1.1.5 Management of Acute Respiratory Failure 8
1.1.5.1 Oxygen supplementation 8
1.1.5.2 Mechanical ventilation 10
1.1.5.3 Continuous Positive Airway Pressure 11
1.1.5.4 Definition of Non-Invasive Respiratory Support 11
1.1.6 Predictors of NIRS failure 12
1.1.6.1 Predictors of response to HFNC 12
1.1.6.2 Predictors of response to NIMV 12
1.2 Machine Learning 14
1.2.1 Definition of Machine Learning 14
1.2.1.1 Machine Learning algorithms 15
1.2.1.2 Machine Learning models interpretation 16
1.2.2 eXtreme Gradient Boosting 16
1.2.3 Tabular Prior-data Fitted Network 17
1.2.4 Clinical applications of Machine Learning 18
1.2.5 Machine Learning in prediction of NIRS failure: state of the art 18

1.3 Frailty 20
1.3.1 Definition of frailty 20
1.3.2 Frailty scores 20
1.3.3 Frailty and age 22
1.3.4 Frailty in prediction of NIRS failure: state of the art 22
1.4 Aims of the study 23
2 Materials and Methods 24
2.1 Study setting and population 24
2.2 Data collection 25
2.2.1 Clinical data 25
2.2.2 NIRS failure definition 25
2.2.3 Frailty assessment 25
2.2.4 Machine Learning models 25
2.3 Analysis plan 26
3 Results 28
3.1 Demographic and clinical data 28
3.2 Predictive power of frailty and age 29
3.3 Effect of frailty on TabPFN model performance 29
3.4 Effect of frailty on XGBoost model performance 33
3.5 Effect of frailty on TabPFN model performance in HFNC-only subgroup 35
4 Discussion 37
4.1 Importance of explanatory models: age and frailty 37
4.2 Impact of frailty on NIRS failure 37
4.2.1 Failure definition 37
4.2.2 Analysis plan differences and heterogeneity 38
4.2.3 Timeframe and late failure 39
4.2.4 Improvement in an integrated model 39
4.3 Comparison of clinical data contributions 39
4.4 Stronger performance in the HFNC-only subgroup 39
4.5 Points of strength 40
4.6 Limitations 41
5 Conclusions 43
6 References 44
7 Acknowledgments 58

1 Introduction

1.1 Acute respiratory failure

1.1.1 Definition of Acute Respiratory Failure

Acute respiratory failure (ARF) is defined as an acute and progressive condition of failure in oxygenation of arterial blood or carbon dioxide elimination [1,2,3] that develops within hours, days or up to a month after the onset of the causing event with no chance for the organism to develop a compensative mechanism [4,5]. It is conventionally recognised by an arterial oxygen tension (PaO_2) < 60 mmHg and/or an arterial carbon dioxide tension ($PaCO_2$) > 45 mmHg [1], or, when the patient is receiving some form of external respiratory aid with additional oxygen support expressed by a fraction of inhaled oxygen (FiO_2) as the concentration of oxygen in the inhaled gas mixture, a PaO_2/FiO_2 ratio ≤ 300 [4]. A low PaO_2 is a condition usually referred to as hypoxia or hypoxemia. PaO_2 should be, in normal conditions, inferior to the oxygen tension in the alveolar air (PAO_2) only to then become identical after the transit time: the difference between the two [$D(A\text{-}a)O_2$] becomes pathological when it becomes more than 5 mmHg [1], though this threshold also depends on other elements such as the patient's age [6]. The diagnosis, generally guided by these laboratory results, strongly depends upon the clinical history and conditions of the patient as well [5].

1.1.2 Pathophysiological classification of Acute Respiratory Failure

ARF can be classified according to the pathophysiological mechanism in two main categories: hypoxemic ARF (type 1 ARF) and hypoxemic-hypercapnic ARF (type 2 ARF) [1]. A rather quick and superficial view of the respiratory system sees it as composed of mainly a lung part and a pump part: a failure in the former causes hypoxemic ARF, a failure in the latter causes hypoxemic-hypercapnic – also known as simply hypercapnic – ARF [9].

1.1.2.1 Type 1 ARF

This form of ARF is characterized by a low level of PaO_2 and a relatively normal level of $PaCO_2$. Four main underlying mechanisms can be responsible for this condition: [10,11]

1. Ventilation/Perfusion ($\dot{V}/\dot{Q}$) mismatch: this is the most common cause [11]. Conditions of either decreased ventilation or perfusion to specific areas of the lung are causes of hypoxemia [1]. $\dot{V}/\dot{Q}$ mismatch may also cause hypercapnia: though oxygen supplementation resolves the hypoxemia through increasing PAO_2 and therefore increasing PaO_2 as well, even if it maintains the high $D(A\text{-}a)O_2$, it can also impede the hypoxemic vasoconstriction that limits the $\dot{V}/\dot{Q}$ mismatch; by limiting this, oxygen supplementation causes the $\dot{V}/\dot{Q}$ ratio to decrease, limiting the possibility of CO_2 to be eliminated and therefore causing hypercapnia, also due to the relatively increased perfusion [2,10]. Decreased ventilation to normally perfused areas occurs in airway obstructions such as obstructive diseases like acute exacerbation of Chronic Obstructive Pulmonary Disease (COPD) or asthmatic attacks, pulmonary oedema and alveolar collapse such as in atelectasis [2]; the opposite condition, with low perfusion and normal ventilation, is typical of pulmonary embolism [10]. The latter condition is characterized not only by early hypoxemia (according to this mechanism), but also by hypocapnia due to tachypnoea [12].
2. Increased shunt: this means blood passes through lung areas that are not ventilated because of acute reductions in lung tissue that can be reached by air [10], like in atelectasis, pneumonia or ARDS. Deoxygenated blood passes by non-ventilated alveoli, maintaining its venous characteristics and mixing in the newly oxygenated blood coming from normally aerated and perfused areas of the lung [5,10].
3. Diffusion reduction. To better understand this cause, it is fundamental to consider the Fick Law that rules gas diffusion: $\dot{V} \propto \frac{A \cdot D \cdot \Delta P}{T}$. $\dot{V}$ represents the rate of diffusion namely the amount of gas that passes through the barrier per unit of time, A is the surface area, D is the diffusion constant which is specific per gas measuring how easily it can diffuse through the specific membrane, ΔP is the pressure gradient through the membrane and T is the thickness that is the distance the gas must travel [7].

   This condition may be due to:

- Decreased A: it may happen in emphysema due to the destruction of alveolar walls and therefore internal alveolar surface, or surgical lung resections [5,10].
- Decreased ΔP: a low inspiratory air pressure can be responsible for decreasing the gradient. PAO_2 also depends, however, on FiO_2 according to the alveolar gas equation $PAO_2 = (FiO_2 \cdot 713\ mmHg) - ({}^{PaCO_2}/_{0.8})$, therefore a decrease in FiO_2, as may happen in high altitudes, can also decrease ΔP [2,5,10].

- Increased T: this may be seen in Interstitial Lung Diseases (ILDs), pulmonary oedema or inflammatory conditions such as ARDS, pneumonia or sarcoidosis which cause fluid accumulation or cellular infiltration, thickening the barrier [2,5,10].
- Decreased exposition time available for blood to receive new oxygen because of increased blood flow rate, as it may happen during intense physical exercise in which there is both an increased need for blood oxygenation and a quicker intrapulmonary blood flow reducing the time window during which blood can be oxygenated [1,5,10].

4. Alveolar hypoventilation: this may cause not only hypoxemia but also hypercapnia and therefore type 2 ARF, as can be seen below (*see type 2 ARF, 1.1.2.2*) [5,10].

1.1.2.2 Type 2 ARF

This form of ARF is characterized by both hypoxemia and hypercapnia. The former can be caused by the conditions listed above, the latter by other causes, some of which match with the causes of hypoxemia. The main causes of hypercapnia are: [10,11]

1. Hypoventilation, which also is responsible for hypoxemia. $PaCO_2$ is linked to $\dot{V}CO_2$ (production of CO_2 per unit of time) and $\dot{V}A$ (alveolar ventilation per unit of time) by the following respiratory equation: $PaCO_2 = \frac{\dot{V}CO_2 \cdot k}{\dot{V}A}$ with k as a constant of proportionality. Therefore, hypoventilation is also responsible for hypercapnia [1,5]. Hypoventilation is mainly a matter of pump failure, behind which three main causes sit [1,5,10,13]:
   - Reduced activity in the respiratory centre controlling respiratory muscles due to anaesthesia, drug overdose or damage of the medulla. This respiratory center receivs and integrates several different feedbacks into a respiratory drive. Some causes for impaired activity of this centre are sedative medications such as narcotics, benzodiazepines, tricyclic antidepressants; stroke; hypothermia [1,5,10,13].
   - Mechanical defects in the muscles or chest wall, e.g. flail chest, Guillan-Barrè syndrome, myasthenia gravis, amyotrophic lateral sclerosis, multiple sclerosis. Such conditions can also be mimicked in major hyperinflation states that flat the diaphragm reducing its moving capacity, such as acute asthmatic attack [1,5,10,13].
   - Muscular fatigue under excessive inspiratory load. This condition comes with an imbalance between energy generation from the muscles and energy payload demanded to the muscles. A decrease in the maximum force that can be generated by the muscle may be due to atrophy, as in prolonged mechanical ventilation, immaturity,

neuromuscular diseases, peripheric nerve damage, muscle blood flow reduction, electrolytes alterations such as calcium serum levels, PaO_2 and inability of the muscles to extract oxygen and nutrients from the blood flow [9]. A muscular inability to extract oxygen and nutrients from the blood flow may be due to conditions such as sepsis or cyanide poisoning; similar situations are seen in low levels of energy stores and glycogen depletion, as in extreme inanition. On the other hand, higher payload can be caused by higher airway resistance, lower thoracic compliance and hyperinflation determining higher threshold load – as already mentioned above, such situations are particularly relevant in the onset of type 2 ARF – [5,9,10,13].

The increase in CO_2 within the alveolar space will cause it to take space from O_2, reducing PAO_2 and therefore PaO_2 as well, causing hypoxemia, too [5,10].

2. Increased CO2 production, like in hyperthermia and during muscular exercise such as inspiratory effort – especially inspiratory resistive breathing –, shivering or increased muscle tone like in tetanus. Other conditions may be sepsis or thyrotoxicosis, since they increase metabolic rate. Additionally, environmental exposure to carbon dioxide can be witnessed in areas close to volcanoes or geothermally active locations. These are rare causes of hypercapnia that barely contribute to type 2 ARF [1,5,10,13].
3. Increased dead space. In every healthy individual there is a normal quantity of dead space (V_D), a volume of air that does not contribute to gas exchange: this is, in normal conditions, only equivalent to the anatomical dead space, which is correspondent to upper airways. The air amount that, at the end of end of the previous expiration, remaines within the anatomical dead space mixes with the newly inhaled air and returns to the alveoli, increasing the alveolar volume (V_A); however, part of the newly inhaled air remains in the upper airways, increasing the anatomical dead space. Therefore, at each breathing cycle, the total amount of air within the respiratory system which is the tidal volume $V_T = V_D + V_A$, from which we assume $V_A=V_T-V_D$. In normal conditions, V_D is about 30% of V_T, which is the percentage of anatomical dead space in a healthy individual. V_D is altered in several pathological conditions [7,20]. Bronchoconstriction causes V_D to decrease, while bronchodilation and tissue traction, such as in emphysema, cause it to increase [5,7,10,13]. It is fundamental to underline that dead space is a condition of ventilation without perfusion, therefore other causes may be pulmonary capillary compression in Positive Pressure Ventilation (PPV) or destruction of pulmonary capillaries like in ILDs. Massive pulmonary embolism can also significantly increase dead space [13]. Under the assumption of an increase in V_D,

since alveolar ventilation $\dot{V}_A = V_A \cdot RR$ (respiratory rate), and since $V_A = V_T - V_D$, we prove V_A and therefore $\dot{V}_A$ to decrease; hence, by the respiratory equation as shown above, $PaCO_2$ increases into a condition of hypercapnia [2,7].

4. Decreased tidal volume (V_T), by the same equation $V_A = V_T - V_D$, causes alveolar hypoventilation and, by the respiratory equation, $PaCO_2$ to increase. It is the volume of air that enters the respiratory airways for each respiratory cycle. Smaller inhalations cause V_T to decrease [5,7,10].

1.1.2.3 Oxygen-induced hypercapnia

Oxygen-induced hypercapnia in COPD patients receiving supplemental oxygen beyond the ideal peripheral capillary oxygen saturation (SpO_2) target of 89-93% [14] is driven primarily by $\dot{V}/\dot{Q}$ mismatch and the Haldane Effect. As to the former, excess oxygen maintains high levels of PAO_2 despite decreased ventilation, preventing perfusion from dropping accordingly to ventilation, leading therefore to a decrease in the $\dot{V}/\dot{Q}$ ratio. The Haldane effect, on the other hand, describes how an increased PaO_2 reduces the affinity of haemoglobin for carbon dioxide. In COPD patients, supplemental oxygen increases oxygenated haemoglobin — which is unable to efficiently bind CO_2 — causing a rightward shift in the CO_2 dissociation curve. Consequently, this leads to an increase in $PaCO_2$ and the onset of hypercapnia [7,8].

1.1.3 Clinical presentation of Acute Respiratory Failure

1.1.3.1 Clinical presentation of hypoxia in type 1 ARF

Main symptomatology includes dyspnoea, a broad term that includes air hunger and chest tightness. This is mainly a self-reported symptom, and though it is possible to quantify it through well-established scales such as the Borg scale and the visual analog scale [10], it should be assessed alongside objectifiable signs such as central cyanosis – a bluish appearance of skin and mucous membranes due to deoxygenated haemoglobin being over the serum level of 5 g/L and that also involves lips and tongue as opposed to peripheral cyanosis due to local vasoconstriction –, tachypnoea – defined as a state of a RR $\geq$ 22 breathes/min – or use of accessory respiratory muscles – which include the sternocleidomastoid and the internal intercostal muscles [15]. We also find signs of adrenergic activation: tachycardia, hypertension or cold sweating. Notwithstanding, particularly severe and rapidly developing cases may manifest through opposite signs such as bradycardia and hypotension, hinting at an imminent

cardio-respiratory arrest. There may also be neurological disorders such as headache, attention deficit, insomnia with daytime lethargy, confusion, anxiety, aggressivity, tremors or psychomotor agitation to delirium; a condition of hypoxemic-ischemic encephalopathy may take place when the previously mentioned neurological conditions are left untreated and develop into seizures, loss of consciousness or comatose state [5,9,10].

1.1.3.2 Clinical presentation of hypercapnia in type 2 ARF

Signs and symptoms are mostly the same as in hypoxemic ARF, even though the neurological alterations tend to be others, secondary to hypercapnia. These belong to a spectrum, the severity of whose symptoms depend on the $PaCO_2$ levels. With mild hypercapnia the patient may report non-specific headache to somnolence, alongside mild dyspnea and tachypnea which are in common with type 1 ARF. Onto higher levels of hypercapnia, the patient becomes disoriented, confused, delirious while the RR slows down and the patient develops bradypnea. In fact, acute hypercapnia initially increases the respiratory drive but, over time, reduces it. The final stages include severe loss of consciousness and coma: this condition is known as hypercapnic narcosis [5,10,13].

1.1.4 Diagnosis of Acute Respiratory Failure

If the patient shows any of the typical clinical features as previously explained, a thorough examination should be assessed [2,5,10].

1.1.4.1 Pulse oximetry

Pulse oximetry allows for measurement of peripheral capillary oxygen saturation of blood (SpO_2). The normal range should be considered 96-100%. Main limits to the use of pulse oximetry are nail polish, poor peripheral perfusion, hypothermia, vasoconstriction, methemoglobinemia, carboxyhaemoglobinemia.

1.1.4.2 Arterial Blood Gas analysis

Arterial Blood Gas (ABG) analysis is the cornerstone to identify ARF. This test allows for a measurement of some fundamental parameters: measured pH, PaO_2, $PaCO_2$, and the consequent calculation of the PaO_2/FiO_2 ratio. The first three reflect the respiratory status of the patient, though PaO_2 should always be evaluated considering whether the patient is receiving supplemental oxygen using therefore the PaO_2/FiO_2 ratio [5,16]. The pH is a way to express the hydrogen ion concentration using the following

equation: $pH = -\log([H^+])$, with the concentration expressed in moles per Liter. Its normal range is 7.35-7.45: below it we find a condition of acidosis, while above it is a condition of alkalosis [17]. Acidosis can be the expression of, among other conditions, type 2 ARF. Whenever the clinical presentation of a patient provides reason to suspect ARF, ABG analysis provides enough data to assess it. The differentiation between type 1 and 2 ARF is provided by $PaCO_2$ [5,16].

1.1.4.3 Laboratory exams

The laboratory evaluations, as any other medical exam, should be guided by clinical suspect and a balance of pros and cons [18]. A complete blood count (CBC) is usually asked for all patients as it has weak cons and provides basic information on oxygen carrying capacity; white blood count (WBC) can help in case an infection was suspected [19], alongside blood or sputum culture, rapid viral panel by polymerase chain reaction (PCR) or urine antigens [2].

1.1.4.4 Chest radiography

Chest radiography is the usual primary imaging technique [19]. It can assess consolidation pleural fluid, hyperinflation, pneumothorax (PNX). It is rather economical and easily-performed [18,19], though it has low sensitivity in acute settings for small lesions, low specificity in its findings and low localization capacity in its bidimensionality, not to mention possible technical errors which have recently been overcome with digital machinery [5,18].

1.1.4.5 Ultrasound

Ultrasound has been proven to have a higher sensitivity in detecting smaller near-pleural phenomena, pleural effusion and PNX than chest radiography, and has higher negative predictive value for heart failure. This technique has grown in the past 15 years, becoming an essential of the Point-Of-Care Ultrasound (POCUS) approach in bed-side context [2,5,19,22]. The main goal is looking for the “B-lines”, which are artifacts manifesting as hyperechogenic, perpendicular lines generating from the pleura and representing interstitial filling, thickening of interlobular and intralobular septa, interstitial pulmonary oedema and pulmonary fibrosis, when at least three of them are observed per each intercostal space; but may also indicate atelectasis or consolidations [5,23]. The importance of lung POCUS has led to the creation of the Bed-side Lung Ultrasound in Emergency (BLUE) protocol for ARF assessment in the

Emergency Department (ED) [20,21,23]. Furthermore, ultrasound should also be used in cardiac terms: rapid bed-side echocardiography can identify cardiac reasons for ARF such as heart failure, pericardial effusion or PE [2]. Venous ultrasound is also to be included, as it may help locate venous thrombi that may help in the context of suspect for PE. Ultrasonography has its main advantage in the absence of radiation, convenient speed of execution and usability in unstable patients; it is however operator-dependant and may suffer in its imaging results from subcutaneous fat and air [18].

1.1.4.6 Chest CT

Chest Computed Tomography (CT) scan represents a use of X-rays that is overall more intense than in a simple radiography causing significant radiation exposure, but performs better in showing, identifying and locating lesions. CT scan proves to be a gamechanger in the identification of infective or inflammatory conditions, ILD and, through the obtaining of images in the arterial phase of the injection of contrast fluid, PE [2]. The use of contrast fluid, however, exposes to a risk for renal failure in nephropathic patients and may be responsible for allergic reactions [5,18].

1.1.5 Management of Acute Respiratory Failure

Management of ARF must exclusively mean to be a stabilizing course of treatment alongside the diagnosis and treatment for the aetiological cause of it [10]: aetiological treatment and the most fundamental form of ARF treatment which is oxygen supplementation are the basis of the escalation therapeutic strategy [3]. The initial approach should include an airway assessment in a life-saving context: relevant compromission of the airways can be ruled out in a patient with valid verbal response, while stridor, gurgling, intolerance to secretions, soft or hoarse voice should be a relevant cause for concern. Chin lift and jaw thrust are to be considered and in case performed in such cases, or if blood or secretions are the responsible factors suction should be performed [2]. Hypoxemia and hypercapnia are two different conditions that respond to different courses of treatment [5].

1.1.5.1 Oxygen supplementation

This is the cornerstone of management for type 1 ARF. It consists of giving the patient the chance to breath in an air mixture with a higher FiO_2 than the 21% that is usually found in open air. The goal is to reach SpO_2 levels above 94% in non-hypercapnic

patients and between 88-92% in hypercapnic patients for the reasons listed above (*see 1.1.2.3.: oxygen-induced hypercapnia*). The main possibilities for supplementation are classified according to flow speed [5].

Low-flow oxygen systems do not allow for precise and constant regulation of flow rate and mostly include [10]:

- Low-flow nasal cannulas (LFNC): these provide supplemental oxygen exclusively inside the nose of the patients, being therefore efficient only in a compliant patient who does not depend mostly on oral ventilation.
- Simple face mask: this manages to cover both nose and mouth, overcoming the limit of requiring a high level of patient compliance in the way of breathing [5]. It has side exhalation ports for letting CO_2 exit the system [2,10].
- Face mask with reservoir: the same facemask used in the simple kind can be connected to a plastic bag of about 600-1000 mL through a one-way valve to prevent rebreathing [10].

High-flow supplementation consists of mainly two devices [10]:

- Venturi mask: this is the most efficient and safe way to administer oxygen [5]. It takes advantage of the Bernoulli principle and the Venturi law: according to the former, a decrease in diameter of a tube brings to an increase in flow rate; according to the latter, the consequence of an increase in speed is a decrease in inside pressure [23]. This causes air to enter from the outside, which is at open air pressure, through specifically designed holes. Setting a specific flow rate, a specific diameter is chosen at which, because of what was said before, is assigned an estimated FiO_2. Each value can be obtained using diversely designed nozzles, identified by specific colours. These masks provide a FiO_2 between 24-60% up to 60 L/min [10,23].
- High-flow nasal cannula (HFNC): this device can not only provide flows between 30-60 L/min and a FiO_2 up to near 100% but also heat up to 37 °C and humidify air up to 100% of saturation, increasing comfort [2,5]. This system also creates a slight Positive End-Expiratory Pressure (PEEP) reducing inspiratory muscle effort. Furthermore, HFNC reduces dead space in the upper airways. Summing up all these factors, HFNC has proven to be efficient in reducing adverse outcomes in hypoxemic ARF [10].
  Moreover, Non-Invasive Mechanical Ventilation (NIMV) and Continuous Positive Pressure Airway (CPAP) can also be a choice in type 1 ARF, as well as in type 2 as

mentioned further on. NIMV has been proven to improve dypnea, gas exchange and muscular effort in hypoxemic ARF, while CPAP alone can improve oxygenation but not muscular effort [24]. However, other studies have disproved significant recommendations for one strategy over another in most cases of hypoxemic ARF [19].

1.1.5.2 Mechanical ventilation

Mechanical Ventilation (MV) in Positive Pressure Ventilation (PPV) is a fundamental in the management of hypercapnic ARF because this form of ARF is an expression of pump failure rather than lung failure [5]. Not all conditions necessarily benefit from these interventions such as NIMV: it is for example indicated in exacerbations of COPD with respiratory acidosis, but its superiority on standard care has not been proven in, for example, acute asthma attacks [19]. MV has three main goals: 1) improvement of gas exchange: it can improve PAO_2 and therefore PaO_2; 2) reduction of inspiratory effort and muscular fatigue, as it causes hypoventilation (*see the causes for type 2 ARF, 1.1.2.2*) which requires external MV to be managed; 3) gaining time to allow spontaneous physiological mechanisms or external treatment to work on the aetiological cause [5]. Invasive MV (IMV) makes use of oro-tracheal intubation (OTI) or tracheostomy and is reserved to the most severe cases [5]. NIMV, on the other hand, is a viable option in non- or semi-intensive settings such as Emergency Department (ED) or ordinary units. Its main interfaces are nasal, oronasal or facial mask, or the helmet [23]. Nasal mask may cause skin irritation but is usually preferred in chronic need for NIMV; oronasal mask is better at preventing air leaks than the nasal mask but interferes with speech, alimentation and may be responsible for claustrophobia and is therefore unrecommended in chronic use. Facial masks have been recently developed in improving comfort and performance. The helmet is a transparent hood that envelopes the whole head: it is generally well-tolerated since it allows speech and alimentation in a way that oronasal and facial masks cannot and has better comfort than a nasal mask [24]. However, compared to an oronasal mask, it has a higher risk of rebreathing at lower flow rates, which therefore always need to be as high as possible, and is more expensive [23].

The main setting for NIMV is Pressure Support Ventilation (PSV). This is a form of ventilation in which the patient autonomously breathes, and on their spontaneous inspiratory trigger a Support Pressure (SP) is set to be administered during inspiration, above a PEEP that allows for a constant positive pressure environment – the

advantages of PEEP are furtherly described in the explanation of CPAP (*see 1.1.5.*3). PSV is flow-cycled: inspiration is followed by exhalation when inspiratory flow rate reaches a value of typically 25% of the peak inspiratory flow rate [25]. Stand-alone PSV is at higher risk for dyssynchrony, which is a major cause for NIMV intolerance. Air leaks may cause this scenario, as the required flow reduction to trigger the expiration valve may be reached later, prolonging inspiration. In such cases, it is useful to set a maximal inspiratory time [24].

With the add-on of the SP above PEEP an Inspiratory Positive Airway Pressure (IPAP) is reached: it is possible to imply that IPAP = PEEP + SP. By setting desired values of PEEP and SP, it is possible to obtain a satisfying Pressure Support Ventilation (PSV) that acts upon spontaneously triggered inspiration acts; by setting EPAP and IPAP values, a BiLevel Positive Airway Pressure (BiPAP) ventilation is obtained [23].

1.1.5.3 Continuous Positive Airway Pressure

It is fundamental to distinguish Continuous Positive Airway Pressure (CPAP) mode from any form of NIMV as CPAP is not ventilation but rather a delivery of positive pressure at a constant value throughout the whole spontaneous ventilatory act with no inspiratory aid. Its main goal is to increase functional residual capacity and to recruit otherwise collapsed and underventilated alveoli. This allows for a decrease in right-to-left intrapulmonary shunt. Moreover, CPAP reduces the inspiratory threshold, reducing therefore inspiratory work and fatigue [24]. CPAP consists therefore in choosing a right value of PEEP: this should not be either too low because it may not be effective, nor too high as it may cause an increase in alveolar dead space and a reduction in compliance, hence an increase in respiratory workload [23].

1.1.5.4 Definition of Non-Invasive Respiratory Support

All respiratory support aids that range from HFNC to NIMV and CPAP can be summarized as forms of Non-Invasive Respiratory Support (NIRS). These forms of respiratory support have different physiological effects, as HFNC has little to no impact on mechanical ventilation whereas NIMV does; however they all share in common the fact that they act from the physical exteriority of the patient with interfaces that do not go inside the respiratory system of the patient [35].

1.1.6 Predictors of NIRS failure

Since ARF is characterized by a rather high risk for mortality, ability to predict the outcome under various management regimens is essential. Delayed intubation, prolonged IMV as well as unnecessary stay in Intensive Care Units (ICUs) have been proven to increase mortality [27]. The most common targets used to express ARF management success are mortality or failure with a specific approach and need for escalation; and since the typical situation of interest is severe ARF, the main management approaches of which failure may be useful to predict are HFNC – which is cornerstone for severe hypoxemic ARF before accessing MV – and NIMV – because IMV has few further options at its failure [3,28].

1.1.6.1 Predictors of response to HFNC

According to the typical escalation strategy, HFNC failure brings about need for MV, and delayed access to it causes mortality to increase [3,27]. A specifically designed index known as ROX score has been developed for this purpose. Original ROX score is calculated as $ROX = \frac{SpO_2/FiO_2(\%)}{RR\ (breathes/minute)}$ [29]. Since HR as well was found to be strongly associated to HFNC failure [30], a modified ROX (mROX) score was proposed as $ROX/HR(beats/minute)$ and was found to perform better than original ROX with highly specific cutoffs in differentiating failure and success at intermediate time stages [34]. As a cutoff, values ≥4.88 of the original ROX score after 12 hours of HFNC were first associated to higher chance of success [32], though the following work that led to mROX found that mROX with a cutoff of 6.80 performs better even in terms of specificity than an original ROX score with a cutoff increased at 5.80 and therefore a theoretically increased specificity [31].

1.1.6.2 Predictors of response to NIMV

NIMV failure, which happens in between 25-59% of patients, brings to either death or IMV delayment, which is associated with a dramatic increase in mortality. A different score has been developed known as the Heart rate, Acidosis, Consciousness, Oxygenation and Respiratory rate (HACOR) score, in light of the fact that these parameters – misured in respectively beats per minute, pH, Glasgow Coma Scale (GCS), PaO_2/FiO_2 and breathes per minute – have often been observed to be more often altered in patients who fail NIMV in typical conditions of tachycardia, acidosis, low GCS, hypoxemia and tachypnoea. The points are assigned as follows:

- HR: 0 points if <120 bpm, 1 point if ≥121 bpm
- pH: 0 points if ≥7.35, 1 point if 7.30-7.34, 2 points if 7.25-7.29, 3 points if ≤ 7.25
- GCS: 0 points if 15, 2 points if 13-14, 5 points if 11-12, 10 points if ≤ 10
- PaO_2/FiO_2: 0 points if ≥ 201, 2 points if 176-200, 3 points if 151-175, 4 points if 126-150, 5 points if 101-125, 6 points if ≤100
- RR: 0 points if ≤30 bpm, 1 point if 31-35 bpm, 2 points if 36-40 bpm, 3 points if 41-45 bpm, 4 points if ≥46 bpm.

A cutoff value of 5 points is generally used, and the maximal number of possible points is 25 [33]. The same score was then expanded including baseline data, vital signs, ABG results and the causes for NIMV need being pneumonia, cardiogenic pulmonary oedema, ARDS, immunosuppression and septic shock – evaluated through the calculation of the Sequential Organ Failure Assessment (SOFA) score –. The updated HACOR score has been proven to increase predictive power when compared to the original score, and has been furtherly studied to the point a relevant stratification of low, moderate, high and very high risk was obtained using several cutoff values of 7, 7.5, 10.5, 11 and 14 points [34].

1.2 Machine Learning

1.2.1 Definition of Machine Learning

Machine learning (ML) is an umbrella term that falls under the wide theme of Artificial Intelligence (AI) [35]. It is the result of the application of the concept of Artificial Neural Networks (ANNs) [36,37]. ANNs mimic the human brain with its neurons: they are composed of large numbers of units which can each have either an excitatory and/or an inhibitory activity, and through their connections an input that comes into the system is turned into an output [38]. These units are organized in layers, each of which is a portion of the network that is differentiated from the others due to having a distinctive function. There is an input layer, where a first collection of data takes place and from which data is passed onto the subsequent layers until the output layer is reached and from which conclusions are gathered. All layers between the input and output ones are known as hidden layers, in which the single actions cannot be observed [37-40].

ML is a field of AI that differs from classical coding techniques because the laws and rules that interconnect data are unknown beforehand: data is provided to the machine, which works on the theoretical concept of an ANN, in the form of an input, and subsequently provides an output [35]. It is fundamental to set up a dataset that can train the machine, which can detect a pattern within this training set and infer relationships invisible to human eye and subsequently apply them to other datasets. A wide enough dataset allows for algorithms to perform well enough; on the other hand, smaller datasets require classical techniques such as linear regression or decision-tree methods that follow fixed rules [35,36,43]. A decision tree is represented by a graph in which each node is an event that leads to a fixed number of possible choices represented by the branches that depart from the node [43]. Decision trees are known to be useful in small datasets leading to small trees: if a decision tree becomes large and complex, each node and subtree includes an excessively small number of data to support it leading to the risk of overfitting [44], which consists in an excessive adherence of the ML model to its training dataset reducing therefore its generalizability and ability to make accurate predictions on data other than that in the training set [45].

1.2.1.1 Machine Learning algorithms

Several ML algorithms can be applied, and none of them are universally adequate. These are [36,43]:

- Supervised learning: this type of ML consists in the model learning the relationship between data in the training set and subsequently applying the acquired knowledge in new contexts [37]. The relationship between data is previously assigned by the "teacher" who associated each data point x_i to a correspondent data point y_i [35]. There are three main steps in supervised ML: 1) splitting the overall dataset in a training set, a validation set and a test dataset [36]. The provision of a training set [41] implies there being external assistance [43]. 2) Use of the training and validation datasets to allow the machine to observe and learn the relationship between items [36]. More specifically, the training data provides the machine with the pre-associated data, while the validation dataset selects the best performer out of a number of different models that may have been trained [35]. 3) Use of the test dataset to prove the ability of the machine to provide an output [36].
- Unsupervised learning: in this case, data is provided with no previously labelled relationship between data. The model learns from a dataset with no expected output values [37] and can be considered to be "optimizing itself". The goal is to detect a hidden pattern withing a dataset [41] without a single correct answer [37] even though there may not be such pattern [36]. It leaves the model to its own and allows it to find interesting and potentially meaningful new patterns, which is useful in more complex processing tasks [41,43].
- Semisupervised learning: this is a medium point between supervised and unsupervised learning, with both labelled and unlabelled data within a single dataset. It is a common form of ML in medical contexts, more specifically for imaging purposes. Even if a skilled physician can only label a small subgroup of images, the model can still benefit from it [36,46].
- Reinforcement learning: this common principle behind learning, beyond just AI, means that certain actions are encouraged by specific feedback with a positive meaning, while others are discouraged. In ML, this form of learning is applied mostly when the goal is to find the most optimal way to reach a certain accomplishment. Through memory of past feedback and exploration of new tactics, the model is driven by the reaching of the biggest final reward possible, not necessarily in the shortest time

possible: it works on a long-term strategy. It is an iterative process: quality improves with higher numbers of rounds [46]. It is also considered the closest ML can get to actual human brain functioning, through error and success, though its medical applications are yet to become substantial [36].

1.2.1.2 Machine Learning models interpretation

ML models can be classified in two main categories: "white-box" and "black-box", the formers including linear regression and decision tree and being the more interpretable ones; the latters being the more complex ones, typically unintelligible from the outside. Black-box models often prove to be more accurate and capable of capturing intricate data patterns. It is possible to interpret black-box models nonetheless – though not always in the most faithful ways – with trained models, such as SHapley Additive exPlanations (SHAP) [47]. This specific model provides SHAP values, which show the weighted contribution of each input variable to a common output in an easily interpretable way [48], but it does not demonstrate a causal relationship: SHAP finds correlations between data, which should not be mistaken for causation. Features may correlate with output data through a proxy, confounding elements or even just by coincidence. SHAP is not capable of discriminating across these possible scenarios, and assigns all relationships with the same importance [125,126]. SHAP values are represented in SHAP Summary Plots, which display on the Y-axis the features ranked by importance as represented by means absolute SHAP values, on the X-axis the SHAP value namely the impact on model output. Each dot is the one patient's SHAP value for that feature and is represented in a colour that correlates with the feature value: red is associated to higher feature value, while blue with lower values.

1.2.2 eXtreme Gradient Boosting

"Boosting" is a term referring to a family of algorithms able to strengthen weak learners and reduce bias and variance. A "weak learner" is any classifier that is not definable as a "strong learner" namely a classifier that is well-correlated with the true classification. This principle is based on the idea of putting together several weak learners to create a single strong learner [43]. These more robust models can help overcome the limit of overfitting that affects decision trees (*see the limits of decision trees in 1.2.1*) [49]. eXtreme Gradient Boosting (XGBoost) is a ML algorithm that applies this principle in various contexts including, more recently, medicine, taking

several predictors – created for example for diagnosis, prognosis or treatment selection – and aggregating them into a more accurate and powerful single one [37]. This kind of models have been the gold standard for tabular data ML over the past 20 years [50], though susceptible to data scarcity and out-of-distribution shifts, which are rather common in research [51].

1.2.3 Tabular Prior-data Fitted Network

Tabular data are a specific form of information expression in a two-dimensional visual mode through rows and columns, namely tables [52]. Over the years, several ML models dedicated to tabular data have been developed, more specifically categorized in three types [53]: 1) specialized methods based on their specific architecture being trained from scratch on a single tabular dataset, including tree-mimic and gradient-boosting approaches such as XGBoost, which recognize their main limit in such scattered and hardly interconnectable way [53,54]; 2) transferrable methods make use of models that are pre-trained on different datasets applying the concept of cross-modal fine-tuning: a model can be pre-trained on aspecific, data-rich datasets and then, through automated ML models, secondarily focus on a specific sub-dataset [53,55], as was done for example in a study by Bendavid et al. so as to apply general knowledge previously built on non-COVID-19 patients to a smaller population of COVID-19 patients [56]; 3) general models take the same concept of generalizability without the need for an intermediate fine-tuning phase but rather managing different heterogeneous datasets simultaneously. Among general models is Tabular Prior-data Fitted Network (TabPFN) [53]. TabPFN is a ML model that stands out for its in-context learning capacity on unprocessed tabular data [57,58] without need for task-specific retraining [51], its remarkable applicability to narrow datasets with less than 10.000 samples and its impressively fast training time when compared to standard models such as XGBoost [50,59]. It is considered to be a "one-size-fits-all" ML model that can be used transversally, and it is predicted that the near future of ML consists of a choice between self-made specific ML algorithms and the use of a general TabPFN model that can adapt to specific situations [58]. It has the potential to overcome the dominance of gradient-boosting models for the first time in the past 20 years [50] thanks to its robustness under the weak points of gradient-boosting models [51]. It apparently still shows some downsides nonetheless: it proves to be weak under dense, fine-grained correspondence information [60] and it mostly can only handle a number

of samples under 10.000 [59]. It also is rather susceptible to data drift [50], as can be proved to the poor performance of TabPFN compared to traditional indices when measurement errors are hypothesized [61]. The future of TabPFN includes work to increase its applicability to larger datasets, reducing its susceptibility to data drift and improving fine tuning [50]; moreover, even if the concept of TabPFN is to be a "one-size-fits-all", some subspecialized prior networks should be developed [50,58].

1.2.4 Clinical applications of Machine Learning

ML has several potential applications in clinical context, mostly applied in detection of visual patterns tasks or segmentation of diagnostic images ML can also aid physicians in tasks of classification led to clinical diagnosis, disease predictions, treatment selection and response probability [35,42,62-64]. Robotic surgery may also be AI-assisted in terms of better viewing and procedural modelling [42]. Overall, however, downsides to these applications should also be considered, both in the risk of reduced skills in operators and possibly low output quality due to several factors, such as low quality in the training dataset or limited capacity in considering several comorbidities or overlapping conditions belonging to a same spectrum of disorders [35]. Ethical dilemmas are also to be taken consideration of, in light of the fact patients cannot interact with the decision-making AI machine as they could with human physicians; artificial machines may also make mistakes, the responsibility of which may be difficult to locate. Finally, decisions may be biased in ways that could be considered discriminatory [42].

1.2.5 Machine Learning in prediction of NIRS failure: state of the art

As already stated above, early identification of NIRS failure is of the most crucial importance, and specific scores such as ROX and HACOR have been developed for this purpose (*see 1.1.6 "Predictors of NIRS failure"*). However, they rely on a small number of variables, and though easy in their use and relevant in clinical practice, they may prove to be poorly predictive and difficulty generalizable on large and heterogenous populations; moreover, no universal threshold is agreed upon, and nonetheless its static nature may prove to be inappropriate for such critically dynamic conditions. Several ML models have therefore been developed over recent years to overcome this limit [65,66]. Both HFNC and NIMV have been studied as forms of NIRS in their predictability of failure. As to HFNC, studies such as Temte et al. [68]

and Yu et al. [69] proved superiority of ML models on ROX and mROX scores with both internal and external validation – the latter missing in other studies such as in Wang et al. [67] –. though in exclusively retrospective settings. As to NIMV, some studies have suggested a relevant predictive power [56,70]. However, a recent systematic review and metanalysis by Akgün et al. has gathered information from 14 studies and concluded that poor homogeneity in outcome definition and methodological approach to treatment escalation and poor use of standard parameters such as the Brier score lead to limited space for progress in productive contribution by ML in ARF management [65]. A review by Notaro et al. underlines the necessity for explainability in black-box models using techniques such as SHAP and for correctly timing data collection as early as possible to prevent the risk of having a chronological proximity between the prediction and the outcome. Moreover, it was observed that dynamic approaches that cover longer time frames produce results that are far superior to in studies that only use static observation of baseline data [66].

1.3 Frailty

1.3.1 Definition of frailty

The definition of frailty has not always been universally agreed upon. Several definitions have been given throughout the years, usually with one common feature which is the concept of a multidimensional syndrome of loss of reserves [71]. A generally agreed-upon definition states that frailty is a syndrome due to depletion in physiologic reserves resulting in increased susceptibility to adverse outcomes in minor events [72,73]. The definition of frailty as a syndrome is crucial to identify people at higher risk [74], because people at risk for frailty are characterized by a wide phenotype that includes several conditions such as sarcopenia, anorexia, osteoporosis, fatigue, risk of falls and poor physical health [72] in a dynamic and progressive process that goes through a "pre-frail" stage [75].

1.3.2 Frailty scores

Smaller consensus converges, however, on a common operational definition: two main types have been proposed: one on a physical phenotype and the other on a multidimensional phenotype, the former coming from Fried et al. and the other from the Canadian group of Rockwood et al. working in the Canadian Study of Health and Aging (CSHA) [76]. Fried et al. in 2001 defined the Fried Frailty Score (FFS) and frailty as the recognition, per measurements and/or self-reports, of a phenotype with three or more of the Fried criteria: 1) history of weight loss; 2) weak grip strength; 3) poor energy or self-reported exhaustion; 4) slow walking speed; 5) low physical activity [77]. Rockwood et al., on the other hand, proposed in 2001 as well a different operational approach using a deficit accumulation model, e.g. the CSHA Frailty Index (FI) calculated as the proportion of potential deficits found in an individual on the total number of deficits investigated [78]. The list of items is found in Figure 1.

Appendix 1: List of variables used by the Canadian Study of Health and Aging to construct the 70-item CSHA Frailty Index

- Changes in everyday activities
- Head and neck problems
- Poor muscle tone in neck
- Bradykinesia, facial
- Problems getting dressed
- Problems with bathing
- Problems carrying out personal grooming
- Urinary incontinence
- Toileting problems
- Bulk difficulties
- Rectal problems
- Gastrointestinal problems
- Problems cooking
- Sucking problems
- Problems going out alone
- Impaired mobility
- Musculoskeletal problems
- Bradykinesia of the limbs
- Poor muscle tone in limbs
- Poor limb coordination
- Poor coordination, trunk
- Poor standing posture
- Irregular gait pattern
- Falls
- Mood problems
- Feeling sad, blue, depressed
- History of depressed mood
- Tiredness all the time
- Depression (clinical impression)
- Sleep changes
- Restlessness
- Memory changes
- Short-term memory impairment
- Long-term memory impairment
- Changes in general mental functioning
- Onset of cognitive symptoms
- Clouding or delirium
- Paranoid features
- History relevant to cognitive impairment or loss
- Family history relevant to cognitive impairment or loss
- Impaired vibration
- Tremor at rest
- Postural tremor
- Intention tremor
- History of Parkinson's disease
- Family history of degenerative disease
- Seizures, partial complex
- Seizures, generalized
- Syncope or blackouts
- Headache
- Cerebrovascular problems
- History of stroke
- History of diabetes mellitus
- Arterial hypertension
- Peripheral pulses
- Cardiac problems
- Myocardial infarction
- Arrhythmia
- Congestive heart failure
- Lung problems
- Respiratory problems
- History of thyroid disease
- Thyroid problems
- Skin problems
- Malignant disease
- Breast problems
- Abdominal problems
- Presence of snout reflex
- Presence of the palmomental reflex
- Other medical history

Figure 1: list of variables used by the Canadian Study of Health and Aging to construct the 70-item CSHA Frailty Index [71]

This approach was however recognized as unwieldy by the same authors, who developed in 2005 a more practical score composed of 7 levels that does not derive from a method of counting but rather a clinical judgement. This is the Clinical Frailty Scale (CFS), in which the levels vary as follows in Figure 2. The same group found that the two indices are strongly correlated with one another, and that higher CFS values correlate significantly with higher risk for institutional care or mortality [71].

Box 1: The CSHA Clinical Frailty Scale

1 *Very fit* — robust, active, energetic, well motivated and fit; these people commonly exercise regularly and are in the most fit group for their age

2 *Well* — without active disease, but less fit than people in category 1

3 *Well, with treated comorbid disease* — disease symptoms are well controlled compared with those in category 4

4 *Apparently vulnerable* — although not frankly dependent, these people commonly complain of being "slowed up" or have disease symptoms

5 *Mildly frail* — with limited dependence on others for instrumental activities of daily living

6 *Moderately frail* — help is needed with both instrumental and non-instrumental activities of daily living

7 *Severely frail* — completely dependent on others for the activities of daily living, or terminally ill

Note: CSHA = Canadian Study of Health and Aging.

Figure 2: the CSHA Clinical Frailty Scale [71]

Other scales exist, however, such as the 5-item Fatigue, Resistance, Ambulation, Illnesses and Loss of weight (FRAIL) scale [79], the Edmonton Frail Scale (EFS) [80,81], the Groningen Frailty Indicator (GFI) [82] and a more complex and recent approach as is found in the REgistro POliterapie SIMI (REPOSI) project [83]. Moreover, the World Health Organization has proposed a new concept which is the Intrinsic Capacity (IC) which encompasses mental and physical capacities an individual is provided with in their lifetime [84].

1.3.3 Frailty and age

Indication for geriatric evaluation and exclusion from clinical trials are often based on chronological age, which is however disputably not always an efficient criterion [82]. Frailty correlates better with worse outcomes, and age is merely considered as a proxy for those in light of the increasing median age of worldwide population thanks to increasing well-being for the elderly [76,82]. It still goes without saying that frailty and age correlate, but it is safer to say that frailty is highly prevalent in older age than the opposite [77,82].

1.3.4 Frailty in prediction of NIRS failure: state of the art

Over the past few years, the association of frailty and NIRS or critical care outcome has been the target of a great deal of research in this area. One of, if not the, most used scores is the CFS [73,85-89], though in other instances FI or FFS [73,90] were also used. It has been generally found that higher frailty is associated with higher risk for NIMV [85,89] and HFNC failure [86]. General mortality for critical patients was found to be strongly associated with frailty as well [73,88,91]. It has also been discussed that general associations between frailty and mortality are the strongest at a population level, making frailty a relevant reason to start end-of-life care. Risk stratification should not be based on frailty alone, though frailty is a useful tool to be always considered [92]. It was suggested to make use of frailty assessment when discussing goals of care and advanced care planning with patients [85,89]. It also must be noted that hypercapnic conditions leading to future need for NIMV are more frequently associated with frailty, so much that the failure association is stronger in the COPD population than in the non-COPD one [85].

1.4 Aims of the study

In this study the aim was to determine whether adding the CFS to a ML model improves its predictive performance as to NIRS failure. The primary aim was to determine the incremental predictive value of CFS; a subsequent exploratory analysis was conducted to investigate its performance in specific NIRS subgroup, more specifically the HFNC-only subgroup. To this goal, we made use of feature ablation, a ML technique that demonstrates the relative importance of specific input data by first removing and then including it in the dataset [93]. The difference between the two results is the consequence of how important frailty is in predicting NIRS failure. A pooled cohort from the Respiratory Intensive Care Unit (RICU) of the University Hospital of Modena Policlinico (Italy) was constructed by using several previous studies [94-100], to which data from the NIV Critical Care & Regional Weaning Centre of the University Hospital of North Midlands (UK) was added.

2 Materials and Methods

2.1 Study setting and population

This was a multi-centre retrospective pooled analysis of prospectively collected data from previous studies [94-100] from the RICU of the University Hospital of Modena Policlinico (Italy) and from the NIV Critical Care & Regional Weaning Centre of the University Hospital of North Midlands (UK). 277 patients belong to the Modena combined dataset and 64 to the North Midlands dataset, bringing to a total of 341 patients; these were either immunocompromised or immunocompetent patients. The Modena cohort patients were characterized by type 1 ARF undergoing NIMV in three studies [94,99,100], ARF due to COVID-19 undergoing NIMV in another [95], type 1 ARF undergoing HFNC in other two cases [96,97], type 2 ARF due to COPD exacerbation admitted to HFNC without immediate indications for MV in another study [98]. These patients were overall hospitalized between October 2016 to May 2025 in the RICU of Modena Policlinico. The patients from the North Midlands cohort were all characterized by being affected by type 2 ARF undergoing NIMV.

Inclusion and exclusion criteria slightly varied from study to study. Across all studies patients were >18 years of age. In one study [94] the presence of ARF was defined with a PaO_2/FiO_2 ratio < 200 mmHg or when a decision for NIMV escalation was made by the attending RICU staff, based on clinical conditions regardless of the study. In other two studies [97,98] the PaO_2/FiO_2 ratio was used as a criterion with a threshold < 300 mmHg; in another study [96] the operative definition for ARF patients was a SpO_2 <90% in oxygen supplementation by Venturi mask with an FiO_2 of 50%; the same SpO_2 was used with a threshold of 92% to define ARF patients in one other study [95]. Other two studies [99,100] had ARF defined with SpO_2 values <90% or PaO_2 <60 mmHg at room air, signs of respiratory effort – such as the use of accessory muscles, paradoxical breathing or thoracoabdominal asynchrony –, or tachypnoea with a RR > 25 breaths/min.

In one study [94] the exclusion criteria were COPD, PE, ILD, neuromuscular disease, cardiogenic acute pulmonary oedema or chest-wall deformities, need for immediate OTI because of cardiopulmonary arrest, respiratory arrest, loss of consciousness with respiratory pauses, psychomotor agitation requiring sedation, pH < 7.20, neurological deterioration, massive secretions, hemodynamic instability or major ECG

abnormalities, pregnancy, intolerance to NIMV, hypercapnic respiratory failure of any aetiology with a $PaCO_2$ < 45 mm Hg, home long-term oxygen therapy or denial of informed consent. In another study [97] the same criteria were used, adding concomitant hypercapnia and use of sedative regimens before admission. Hypercapnia, nasal tract alterations ang long-term oxygen regimen were the exclusion criteria in another study [96]. Intolerance to NIRS, pregnancy or inability/refusal to provide informed consent were other criteria for exclusion [98]. For other two studies [99,100] the exclusion criteria were urgent need for OTI when prolonged respiratory pauses, cardiorespiratory arrest, HR<50 bpm with progressive loss of consciousness, pH < 7.15; haemodynamic instability; contraindications to NIMV such as uncontrolled vomiting, copious oral secretions, facial deformities, GCS ≤12 points, PNX or a Do-Not-Intubate (DNI) order.

2.2 Data collection

2.2.1 Clinical data

Clinical data included RR, PaO_2, FiO_2 and PaO2/FiO_2 ratio. These were measured at two-time settings being T0 – baseline values within 6 hours prior to NIRS initiation – and T1 – values recorded 1-2 hours following the start of NIRS.

2.2.2 NIRS failure definition

NIRS failure was ubiquitously defined as treatment escalation – meaning access to NIMV from HFNC or IMV from NIMV – within 24 hours, death within 7 days or physiological failure criteria meaning tachypnoea (RR>20 bpm) or hypoxemia (PaO_2 < 60 mmHg) – even in absence of treatment escalation.

2.2.3 Frailty assessment

Frailty was retrospectively assessed using all pre-admission functional information, medical and nursing records. The investigator was blinded to NIRS outcome during the evaluation and used the CFS criteria. CFS grading was assigned using the most recent pre-morbid functional information available. Missing data were imputed using median imputation prior to model training.

2.2.4 Machine Learning models

The mainly used ML model was TabPFN mostly so as to perform efficiently and with no need for task-specific retraining in a relatively narrow dataset. XGBoost was also

used to repeat calculations and compare results. 10-fold cross-validation was performed both with TabPFN and XGBoost: the dataset is folded randomly in 10 folds – which guaranteed equal distribution of NIRS success and failure outcomes across all folds –, and for each fold 90% of data was used for training purposes and final testing was performed on the remaining 10% [101]. Being both black-box models, SHAP analysis was performed to quantify the contribution of each feature to model predictions by obtaining SHAP values. These were computed using Kernel SHAP, which is a method that through weighted linear regression can compute the importance of each feature [102]. 30 background samples were used, selected via k-means clustering: this is an unsupervised ML algorithm, and it proves useful to substitute the missing variable that is taken away for feature ablation purposes by using a background, unlabelled dataset [103]. As per perturbation theory 1000 perturbation samples were used per test instance, using data from the k-means dataset.

This study was conducted adhering to the TRansparent model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines on transparency in AI-based studies [65,119].

2.3 Analysis plan

Predictive performance was evaluated using several metrics, including Sensitivity, Specificity, Area Under the Curve (AUC) in the Receiver Operating Characteristics (ROC) analysis, Accuracy, Balanced Accuracy and the Brier Score – this last one used as a measurement for precision and model calibration quality.

These variables were accounted for at their Youden-optimal classification threshold. This is a method that relies on the calculation of the Youden index as the maximum difference between Sensitivity and 1-Specificity at all threshold values until a critical threshold value is found at which the Youden index reaches the highest possible value. The critical threshold value is found to be the Youden-optimal classification threshold [124].

Results were reported as mean ± standard deviation. Continuous baseline variables were compared between outcome groups with independent two-sample Student's t-tests – normality of the distributions was verified via visual inspection of plots –; binary variables were compared with Pearson's chi-square tests. Unadjusted associations between individual variables and outcome are reported as odds ratios with 95% confidence intervals from univariable logistic regression.

Where model performance was summarised across repeated cross-validation folds, paired t-tests were used to compare paired metrics across folds. To minimize the risk of overfitting to the relatively small dataset, a minimal set of features (RR, PaO_2, PaO_2/FiO_2 ratio, $PaCO_2$ and mROX) was selected for inclusion in the machine learning models. No missing data were detected. Three feature configurations were evaluated: frailty or age as independent predictors in a logistic regression model; ML models without frailty including clinical data and age; ML models with the aforementioned variables and CFS.

Models were compared using the corrected resampled t-test of Nadeau and Bengio. In ML it is not possible to use normal Student's t-test to compare models because the repeated splitting of data to create different training sets from the same pool brings to a significant overlapping and dependency. Standard Student's t-test assumes independency between the pools, whereas these are actually intertwined. This causes the t-test to underestimate the variance and produce important Type I errors, erroneously concluding a model is better. This test inflates the variance estimate to account for the dependency and counterbalance the underestimate of the variance [122,123].

3 Results

3.1 Demographic and clinical data

In Table 1 a statistical comparison of patient data, both in demographic and clinical terms, is shown; clinical data is presented at both T0 and T1. P-values are also presented regarding correlation of each feature with outcome (success vs failure).

		Type of NIRS		Outcome		p-value of association with outcome
Features	**Overall**	**HFNC**	**NIMV**	**Success**	**Failure**	
Demographic data						
Number of patients	341	192	149	204	137	—
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
CFS	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)	—	—	—
Clinical data at T0						
RR, breaths/min	27.8 ± 7.6	27.0 ± 5.7	28.8 ± 9.4	26.1 ± 6.6	30.4 ± 8.2	<0.001
PaO2, mmHg	66.4 ± 10.4	65.6 ± 10.5	68.1 ± 10.0	66.6 ± 10.0	66.1 ± 11.0	0.678
FiO2, %	51.3 ± 15.9	51.6 ± 13.7	50.9 ± 18.4	48.2 ± 16.1	55.8 ± 14.5	<0.001
P/F ratio, mmHg	134.5 ± 42.5	138.7 ± 44.0	125.0 ± 37.4	142.6 ± 45.5	123.9 ± 35.8	<0.001
pCO2, mmHg	35.9 ± 9.2	36.8 ± 9.8	33.8 ± 7.1	35.9 ± 9.5	35.9 ± 8.8	0.987
mROX index	5.0 ± 2.2	5.4 ± 2.2	4.1 ± 1.7	5.4 ± 2.2	4.4 ± 2.0	<0.001
Clinical data at T1						
RR, breaths/min	24.4 ± 6.2	23.7 ± 5.0	25.3 ± 7.4	22.6 ± 5.4	27.2 ± 6.4	<0.001
PaO2, mmHg	66.5 ± 8.2	66.2 ± 8.9	67.0 ± 6.2	65.9 ± 8.3	67.3 ± 8.0	0.155
FiO2, %	50.1 ± 15.4	50.3 ± 15.7	50.0 ± 15.0	45.9 ± 14.1	56.5 ± 15.2	<0.001
P/F ratio, mmHg	142.4 ± 40.9	144.9 ± 43.4	136.7 ± 34.0	155.5 ± 41.6	125.3 ± 32.9	<0.001
pCO2, mmHg	36.8 ± 8.6	37.9 ± 9.5	34.5 ± 5.7	37.3 ± 7.5	36.2 ± 9.9	0.320
mROX index	6.1 ± 2.7	6.4 ± 2.7	5.2 ± 2.5	6.9 ± 2.6	5.0 ± 2.4	<0.001

Table 1: Statistical comparison of demographic and clinical data at T0 and T1 for NIRS form and outcome. P-values are referred to the association of each feature with either type of outcome (success vs failure)

3.2 Predictive power of frailty and age

A relevant finding, consistent with previous literature [105-107], was that age was not significantly associated with NIRS outcome whereas frailty was. The odds ratio (OR) of the association between age and NIRS outcome was 0.99 – with a p-value of 0.237 – while for frailty the OR was 1.19 and p=0.005. A higher prevalence of frailty was found in the NIRS failure population, with a CFS ≥5 in 43.8% of failures compared with 33.3% of successes, though no statistical significance was reached (p=0.065). This finding was also consistent with previous literature [85,86,88,89,91]. A logistic regression model using frailty alone at its Youden-optimal threshold showed an AUC=0.59±0.07, an Accuracy=0.51±0.06, a Sensitivity of 0.725 but also a Specificity of 0.382. No threshold provided satisfying values of all metrics, underlying a poor overall performance of this standalone predictor.

3.3 Effect of frailty on TabPFN model performance

In Table 2 and in Figure 3, predictive performances of CFS and two different TabPFN models are expressed through the aforementioned metrics. A TabPFN model was trained without the inclusion of CFS, the other with it. The p-values of both the comparison of CFS to a TabPFN-with-CFS model and of the comparison of the two TabPFN models are shown; as to the former the null hypothesis is the equality of the performances, for the latter it was that adding the CFS did not impact the performance.

Metric	**CFS**	**TabPFN without CFS**	**TabPFN with CFS**	**p (TabPFN with CFS vs CFS)**	**p (TabPFN with CFS vs TabPFN without CFS)**
AUC	0.59 ± 0.07	0.77 ± 0.08	0.816 ± 0.07	<0.001	0.001
Accuracy	0.51 ± 0.06	0.69 ± 0.08	0.76 ± 0.07	<0.001	<0.001
Balanced Accuracy	0.56 ± 0.06	0.7 ± 0.07	0.76 ± 0.07	0.484	0.582
Sensitivity	0.725 ± 0.19	0.74 ± 0.08	0.70 ± 0.12	<0.001	0.002
Specificity	0.382 ± 0.15	0.65 ± 0.15	0.8 ± 0.10	<0.001	0.002
Brier score	0.24 ± 0.01	0.19 ± 0.04	0.17 ± 0.03	<0.001	0.003

Table 2: Predictive performance metrics for CFS only (Logistic Regression model), TabPFN model without CFS and TabPFN model with CFS

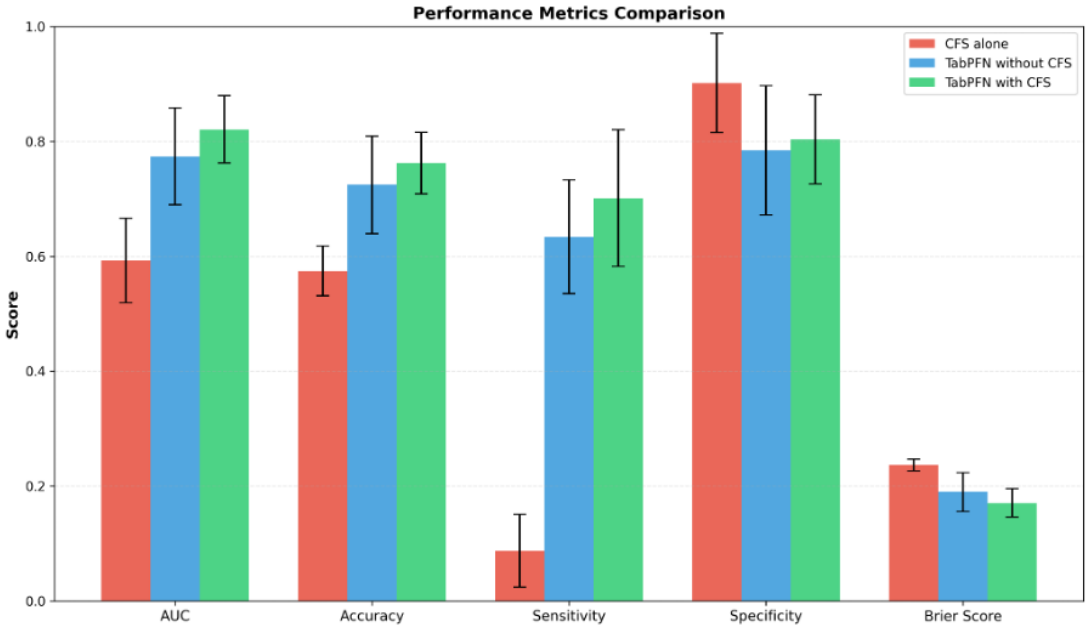


Figure 3: Predictive performance metrics for a logistic regression model using CFS alone, for the TabPFN ML model without CFS and for the TabPFN ML model with CFS

TabPFN proves, under both circumstances, to perform better – with statistical significance – than CFS alone for almost all metrics. Including CFS in the TabPFN model brought to an increase in performance quality for almost all metrics with statistical significance, except only for Balanced Accuracy as to which statistical significance was not obtained. Accuracy increased from 0.69±0.08 to 0.76±0.07 and AUC increased from 0.77±0.08 to 0.82±0.07, with improved probability calibration as can be witnessed by a lower Brier score.

SHAP analysis is shown in Figure 4 and in Figure 5; the former for the TabPFN-without-CFS model and the latter including CFS. RR at T1 (RR_T1) is always the most important feature, but the inclusion of CFS – which significantly increases predictive performance as shown in Table 2 – becomes the second most important feature. Higher CFS scores are associated with higher risk for NIRS failure. Age is poorly linked to NIRS failure – as can be seen in Table 1 – and though of poor influence in the TabPFN-without-CFS model – as in Figure 4 – it acquires high relevance once CFS is included – as can be assessed in Figure 5.

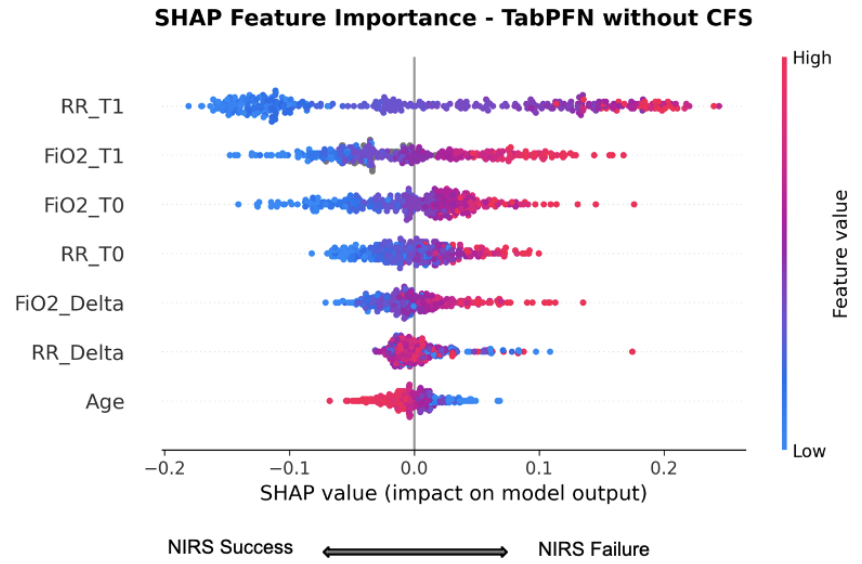


Figure 4: SHAP feature importance plot for the TabPFN ML model without CFS

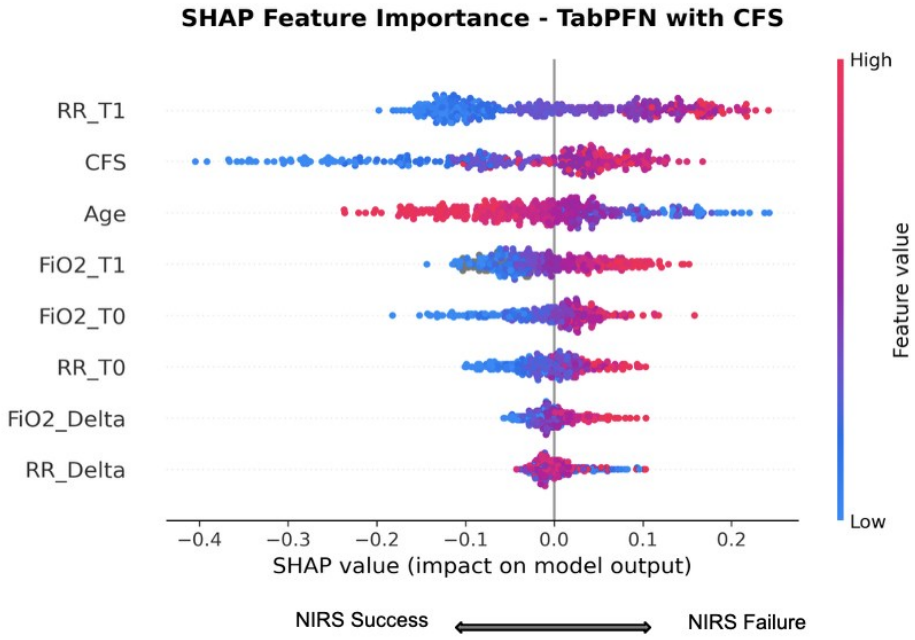


Figure 5: SHAP feature importance plot for the TabPFN ML model with CFS

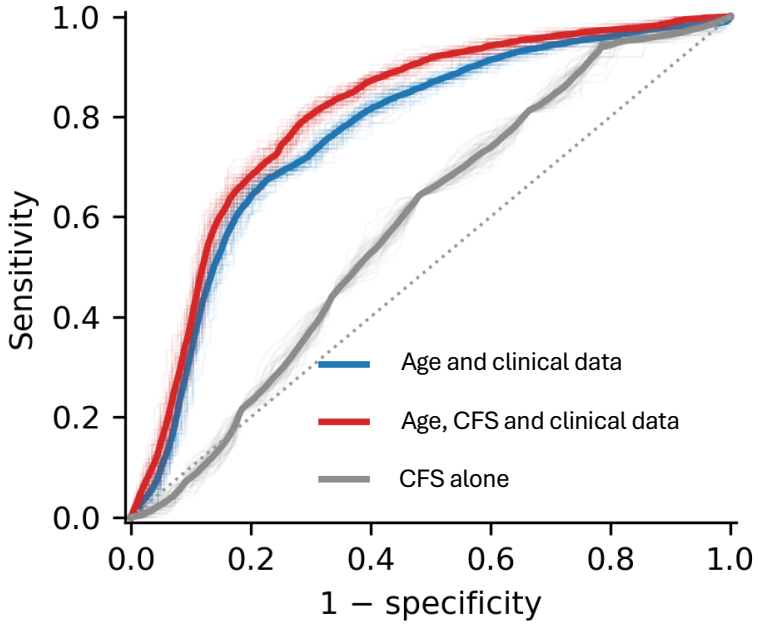


Figure 6: Receiver-Operating Characteristics curve for the TabPFN ML models using only CFS or clinical data (early physiology) with or without CFS as well

3.4 Effect of frailty on XGBoost model performance

Similar results were observed in a repeated feature ablation study using the XGBoost ML model – as shown in Table 3 and in the SHAP analyses in Figure 7 and in Figure 8: a directional coherency can be found with the TabPFN results. Inclusion of frailty was associated with improvement in XGBoost performance with statistical significance for most metrics. CFS showed the second highest SHAP-based feature importance.

Metric	**CFS**	**XGBoost without CFS**	**XGBoost with CFS**	**p (XGBoost with CFS vs CFS)**	**p (XGBoost with CFS vs XGBoost without CFS)**
AUC	0.59 ± 0.07	0.77 ± 0.08	0.81 ± 0.06	<0.001	0.001
Accuracy	0.51 ± 0.06	0.72 ± 0.07	0.76 ± 0.06	<0.001	0.062
Balanced Accuracy	0.56 ± 0.06	0.71 ± 0.06	0.76 ± 0.06	0.765	0.051
Sensitivity	0.75 ± 0.19	0.66 ± 0.09	0.73 ± 0.13	<0.001	0.373
Specificity	0.36 ± 0.15	0.76 ± 0.13	0.78 ± 0.09	<0.001	0.043
Brier score	0.24 ± 0.01	0.2 ± 0.03	0.18 ± 0.03	<0.001	0.002

Table 3: Predictive performance metrics for CFS only (Logistic Regression model), XGBoost model without CFS and XGBoost model with CFS

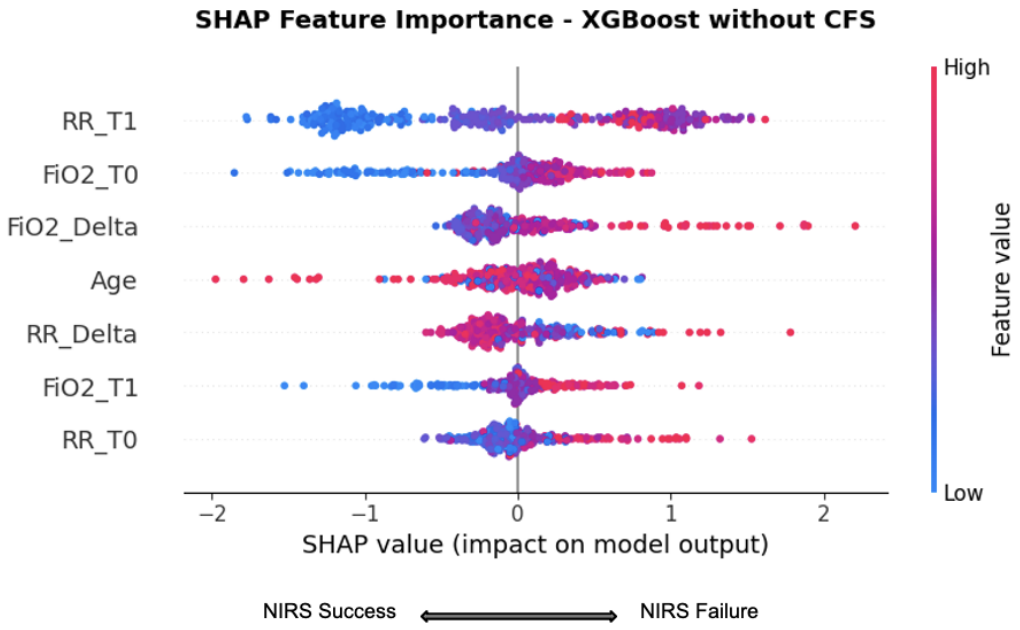


Figure 7: SHAP feature importance plot for the XGBoost ML model without CFS

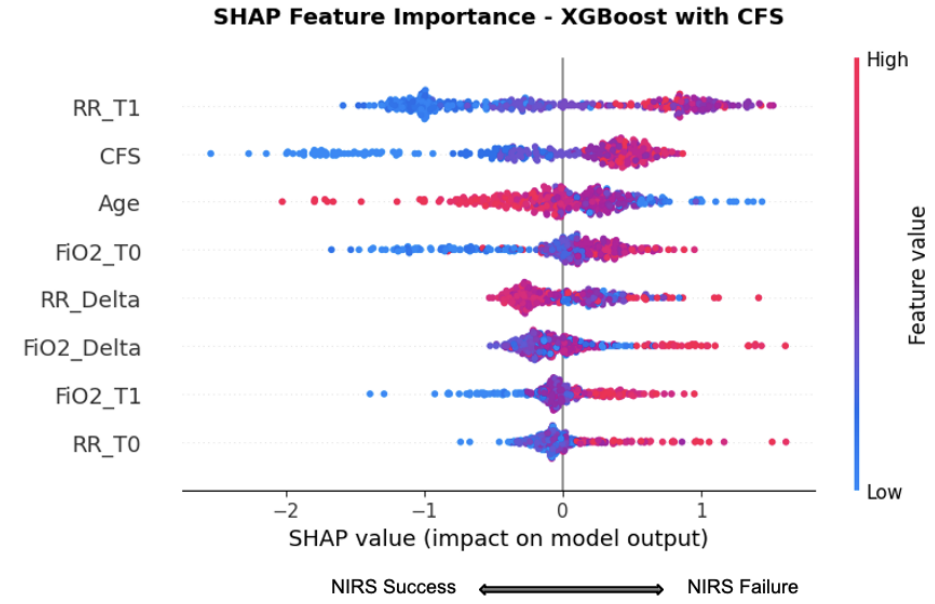


Figure 8: SHAP feature importance plot for the XGBoost ML model with CFS

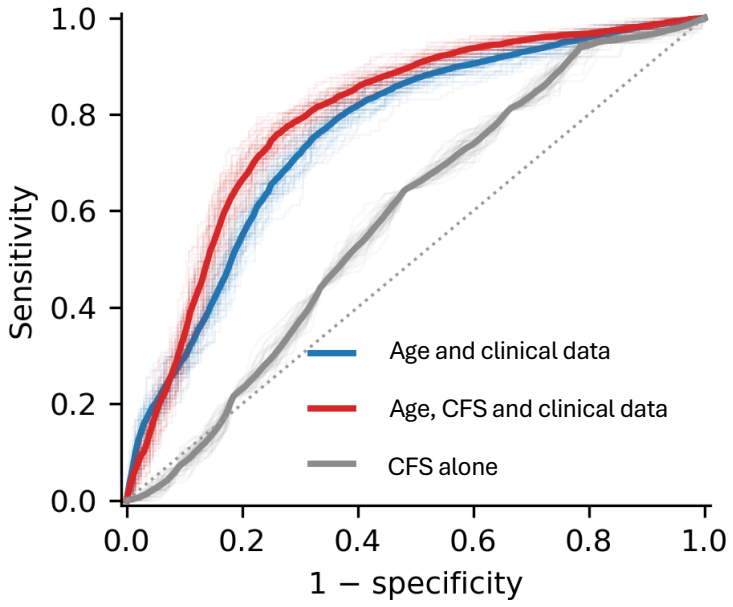


Figure 9: Receiver-Operating Characteristics curve for the XGBoost ML models using only CFS or clinical data (early physiology) with or without CFS as well

3.5 Effect of frailty on TabPFN model performance in HFNC-only subgroup

A similar study was performed with TabPFN on a specific subgroup who were exclusively receiving HFNC (n=192) restricted to the Modena cohort. The performance enhancement through the inclusion of CFS in the TabPFN model was furtherly pronounced, as can be seen in Table 4.

Metric	**CFS**	**TabPFN without CFS**	**TabPFN with CFS**	**p (TabPFN with CFS vs TabPFN without CFS)**
AUC	0.726 ± 0.09	0.889 ± 0.08	0.946 ± 0.035	0.009
Accuracy	0.619 ± 0.122	0.871 ± 0.084	0.907 ± 0.051	0.128
Balanced Accuracy	0.617 ± 0.109	0.853 ± 0.090	0.890 ± 0.065	0.304
Sensitivity	0.617 ± 0.162	0.793 ± 0.143	0.836 ± 0.140	0.178
Specificity	0.618 ± 0.194	0.913 ± 0.095	0.944 ± 0.063	0.168
Brier score	0.202 ± 0.023	0.108 ± 0.042	0.086 ± 0.029	0.013

Table 4: Predictive performance metrics for CFS only (Logistic Regression model), TabPFN model without CFS and TabPFN model with CFS, in the HFNC-only subgroup

The baseline TabPFN model excluding frailty returned an AUC of 0.889 ± 0.080 and an Accuracy of 0.871 ± 0.084. Upon inclusion of the CFS the model achieved an AUC of 0.946 ± 0.035, an Accuracy of 0.907 ± 0.051, and a Balanced Accuracy of 0.890 ± 0.065. SHAP analysis was performed and is shown in Figure 10. CFS showed the highest SHAP-based feature importance.

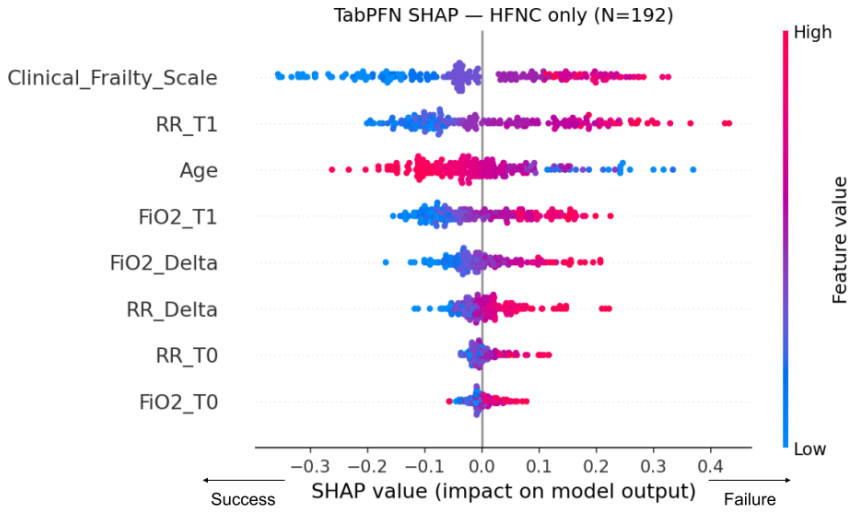


Figure 10: SHAP feature importance plot for the TabPFN ML model with CFS in the HFNC-only subgroup

4 Discussion

4.1 Importance of explanatory models: age and frailty

While chronological age is widely recognized as a mere proxy for frailty [76] and a poor standalone predictor [105-107] – confirming the findings of this study showing age to be contributing minimally to model performance when frailty is omitted – its influence increases significantly when both variables are integrated. Indeed, once frailty is accounted for, age emerges as the third most critical feature in both models (see Figures 5 and 8), with older age associated with lower predicted NIRS failure risk. It is likely that age captures specific differences in underlying disease aetiology. Perhaps older individuals mostly present with acute-on-chronic conditions often of the hypercapnic kind – such as COPD exacerbations or cardiogenic pulmonary oedema – that are highly responsive to NIRS [5,10,23,108], and it has also been observed that NIRS failure was more strongly associated with frailty – which age is related to – in the COPD population than in the non-COPD one [85] ; conversely, younger patients might be more prone to severe hypoxemic conditions with smaller response to NIMV and higher risk of NIRS failure [109]. Thus, frailty and age appear to reflect complementary dimensions of patient risk.

4.2 Impact of frailty on NIRS failure

It has been observed that frailty contributes beyond age and early physiological response variables for early NIRS outcome prediction. Previous literature generally agrees on the presence of a direct relationship between frailty and NIRS or critical care failure [85-92]. However, in this study, frailty alone appears to be poorly associated to NIRS failure. Several explanations can be found for this apparent discrepancy.

4.2.1 Failure definition

The heterogeneity in NIRS failure definition has already been found to be a confounding element [65]. Most studies use, as outcome, 7- to 90-day mortality [85-92], while in this study NIRS failure was defined as treatment escalation within 24 hours, death within 7 days or physiological failure criteria. The timeframe was as narrow as possible so as to adhere to typical clinical decision-making timeframes and reduces variability and redundance [70], but not too close so as to make enough time

between the data collection time and the failure event, making the model truly useful to change therapeutic course of action in a clinically relevant manner [66].

4.2.2 Analysis plan differences and heterogeneity

AUC appears to be used in only one study concerning the matter, in which the AUC value was 0.75 for the CFS alone [88]. In the present study, CFS alone obtained an AUC value of 0.59 ± 0.07. The difference can be due to a smaller number of patients being taken into consideration, which may be one of the limitations of the study; more varied inclusion criteria also created a more heterogeneous cohort, which may make it harder to obtain higher values but also represents more closely the real world [120,121]. However, AUC significantly increased with the use of TabPFN. Indeed, logistic regression seeks additive relationship between frailty and NIRS failure; on the contrary, ML may capture nonlinear relationships that cannot be represented by a simple univariable logistic model. Moreover, it can be argued that a difference is in the interpretation that is given facing certain AUC values: in this case, a satisfying value is found in those higher than the 0.75 that was found in the study by Saxena et al. [88].

It should also be noted that statistical significance was reached for almost all metrics using TabPFN apart from Balanced Accuracy (see Table 2), while for XGBoost statistical significance was only found for AUC, Specificity and Brier Score (see Table 3). Though this surely strengthens the sake of TabPFN, it should not weaken the one of XGBoost. It is relevant indeed to underline the main difference between AUC and the other metrics such as Accuracy. Accuracy is a binary metric that strongly depends on a threshold, which makes it hardly reliable in smaller datasets. Shallow and strong probabilities for the same event happening are all equally represented in the same outcome category. AUC, on the other hand, is the result of constantly changing the threshold and therefore evaluates the quality of the probabilistic distribution on the entire wide dataset. It may therefore prove useful, in the future, to always make use of ROC analysis and use AUC to more accurately depict this kind of relationships.

The Brier score was widely used and trusted in evaluating the value of the prediction models. ARF is a condition in which early identification and therefore sensitivity is crucial, and under-treatment is fare worse than unnecessary treatment; for this kind of situations, a study by Assel et al. argues that the Brier score is not generally clinically

useful as it favours tests with higher specificity and would equally recommend the use of net benefit as it is complementary to the Brier score. Net benefit is a decision-analytic metric that balances true positives against false positives, and it depends on the chosen threshold probability. It decreases when estimated probabilities deviate from actual ones [104].

4.2.3 Timeframe and late failure

Additionally, it has already been stated that frailty is related to later failure rather than that in the timeframe that was hereby considered [107], showing therefore frailty to be only apparently a poor predictor because of the timeframe that was used.

4.2.4 Improvement in an integrated model

All in all, stating that frailty is a poor predictor of failure may reflect its static nature. CFS captures baseline vulnerability. On the other hand, the patient's response to NIRS is actually shown by dynamic physiological responses as represented by changes in RR and FiO_2 from T0 (RR_T0 and FiO_2_T0) to T1 (RR_T1 and FiO_2_T1). It is relevant therefore to stress out that the complex path of a patient's response to NIRS, leading to the definition of what treatment choices may show to be more appropriate with plausible impact on the patient's outcome, depends both on how the patient presents at the beginning of the treatment, as can be assessed not only by clinical data at T0 but also by measuring CFS, and on how the patient responds to the treatment itself, as tested by clinical data evolution at T1.

4.3 Comparison of clinical data contributions

SHAP allowed to identify the most important clinical data in determining the outcome of NIRS in RR_T1 and FiO_2_T1. This confirms what was already stated in previous literature such as Temte et al. [68]. mROX never appeared to be a relevant contributor to outcome prediction as per SHAP analysis.

4.4 Stronger performance in the HFNC-only subgroup

NIMV is known to be a NIRS method that reduces inspiratory effort in a more pronounced way than HFNC, both in hypoxemic [110] and in hypercapnic [111] conditions. NIMV unloads respiratory muscles, though HFNC is still capable of reducing work of breath in a slight way that appears to be significant as opposed to LFNC and face masks [112,113]. Nonetheless, it still goes without saying that NIMV

reduces effort per breath in a greater way [114]. Respiratory impairment has been associated with frailty [115], and an association between frailty and respiratory effort – expressed as respiratory muscle strength in terms of inspiratory and expiratory pressures – has been found as well [116,117], basically because of respiratory sarcopenia, which is the respiratory manifestation of a typical frail phenotype leading to the thinning of diaphragm and intercostal muscles and therefore lower respiratory pressures and flow rates [118]. It might therefore be possible to interpret NIMV not as a NIRS form that is less tolerable than HFNC – as such interpretation would lead to the opposite conclusion namely that frailty should positively correlate with NIMV more than with HFNC – but as more capable to support a frail patient who has a smaller residual functional reserve to autonomously ventilate, though still in accordance with their frailty. Frailty would therefore be a signal for a decline in functional reserve [116-118]. HFNC leaves the patient with a higher workload on their own as to mechanical ventilation: the frailer the patient is, the lower their residual functional capacity is, and therefore the harder it may be for the patient to autonomously ventilate in an HFNC regimen, consequently leading to higher risk for HFNC failure. This could be an explanation for the stronger connection that is seen between frailty and HFNC outcome than with NIMV. It is however not possible to rule out a confounding effect of case-mixing from different datasets and differences in indications for either HFNC or NIMV across clinical settings.

4.5 Points of strength

When compared to most studies regarding NIRS strategies for ARF, some features make this study stand out. The use of ROC-AUC and the geographic diversity of cohorts – though in this case none for external validation purposes – are recommendations by a systematic review by Burak Akgün that were respected in this study; as suggested by the same review, adherence to TRansparent model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines was also respected [65]. At the same time, the use of explanatory models such as SHAP improve the quality of the findings [66]. The limited size of the pooled cohort is surely a downside, but it is possible to state that the use of a parsimonious set of features reduces the risk of overfitting of the model on the training dataset [44]. At the same time, though the heterogeneity in inclusion criteria reduced the chances of obtaining high AUC values, it led to creating a pooled cohort that was as close to the real world as possible

[120,121]. Finally, the consensus between the two distinct ML models reinforces the robustness of the findings.

4.6 Limitations

Several limitations of this work should be noted, which often collide with recommendations by systematic reviews and metanalyses [65,66,119]:

- Clinical data and CFS were all evaluated retrospectively through patient charts. This may expose the findings to selection bias and reducing therefore generalizability and external validation [66].
- Indeed, there was no external validation cohort because all available cases were utilized solely for model training, since the dataset was relatively small despite aggregating data across multiple centres. This negatively affects the generalizability of these findings, which should be interpreted more as hypothesis-generating rather than a true clinical tool, at least at this stage.
- Reduced sample size is a relevant limitation per se beyond its impact on the absence of an external validation cohort.
- A quite broad definition of NIRS failure was necessary, given the heterogeneity across the included cohorts – as the UK cohort is mostly made of hypercapnic patients in NIMV while the Italian one is also consistently inclusive of HFNC cases as well.
- Inclusion and exclusion criteria varied within cohorts in a non-insignificant way.
- Significant heterogeneity also characterized the centres from which patient data was collected, between the RICU of the Modena Policlinico and the NIV Critical Care & Regional Weaning Centre of the University Hospital of North Midlands, in terms of clinical approach and treatment escalation strategies.
- Referring to approach and strategies, it should also be noted that potential treatment limitations and DNI orders may have not always been taken into account, potentially affecting the results.
- The HFNC-only cohort was exclusively based on the Modena cohort, making this evaluation a single-centre analysis, and this nature limits the generalizability of these conclusions.
- Relevant conclusions are drawn based on SHAP results, though SHAP has a non-causal nature, and may take confounding elements into account leading to a possible

mix of correlation with causation (*see 1.2.1.2 "Machine Learning models interpretation"*).

- As to the use of paired t-tests to compare performance metrics across folds, caution should be used due to the inherent data overlap and lack of independence between cross-validation partitions. Consequently, differences in model performance are meant to be an exploratory hypothesis rather than definitive evidence of superiority.
- Adherence to Prediction model Risk Of Bias ASsessment Tool (PROBAST) guidelines was missing, as well as the sharing of open-source code to allow for independent replication, reducing therefore methodological transparency, as well as increasing risk of exposure to risk of bias [65,119].

5 Conclusions

These findings show frailty to be a relevant element in predicting outcome of NIRS in ARF, as it reflects not only baseline patient vulnerability but also how such vulnerability influences clinical decision-making – such as treatment limitations, escalation strategies and NIRS approach. Although the CFS seems to be a poor standalone predictor, its inclusion alongside clinical data consistently enhances predictive performance across two distinct ML models trained on multicentre data. Moreover, even though age is not a relevant predictor per se, it seems to be complementary to frailty in reflecting the consequences of the aetiology, thus it should be accounted for, though not on its own, differently to frequent common clinical practice. Furthermore, frailty showed stronger association with NIRS outcome in a HFNC-only subgroup, suggesting clinical consideration of a possibility to provide NIMV in frail patients in the first place. Pending prospective and external validation, these preliminary insights support further evaluation of frailty as a component of risk-stratification models. Further recommendation for future research in this direction is to include the use of net benefit alongside the Brier Score [104], to always make use of ROC-AUC [65] and to adhere to PROBAST guidelines [65,119].

6 References

1. Roussos, C., & Koutsoukou, A. (2003). Respiratory failure. *European Respiratory Journal*, *22*(47 suppl), 3s–14s. https://doi.org/10.1183/09031936.03.00038503
2. Jacob, S., & Seethala, R. (2026). An approach to airway management and acute respiratory failure. In *Techniques in Vascular and Interventional Radiology*. W.B. Saunders. https://doi.org/10.1016/j.tvir.2026.101113
3. Scala, R., & Heunks, L. (2018). Highlights in acute respiratory failure. In *European Respiratory Review* (Vol. 27, Number 147). European Respiratory Society. https://doi.org/10.1183/16000617.0008-2018
4. Fujishima, S. (2023). Guideline-based management of acute respiratory failure and acute respiratory distress syndrome. In *Journal of Intensive Care* (Vol. 11, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s40560-023-00658-3
5. Oliaro and Loizzi. (2021). *Malattie dell'Apparato Respiratorio*. ISBN: 9788855320542.
6. Hantzidiamantis, P. J., & Amaro, E. (2023). *Physiology, Alveolar to Arterial Oxygen Gradient*. StatPearls .
7. Grassi, F., Negrini, D., & Porro, C. A. (2021). *Fisiologia umana*. ISBN: 9788895033990
8. Abdo, W. F., & Heunks, L. M. A. (2012). Oxygen-induced hypercapnia in COPD: myths and facts. In *Critical Care* (Vol. 16, Number 5). https://doi.org/10.1186/cc11475
9. Oana, S., & Mukherji, J. (2014). Acute and chronic respiratory failure. *Handbook of clinical neurology*, *119*, 273–288. https://doi.org/10.1016/B978-0-7020-4086-3.00019-9
10. Rugarli, C., Cappelli, G., Cappellini, M. D., Cavallo Perin, P., Corazza, G. R., Craxì, A., Crea, F., Fabbri, L. M., Ferraccioli, G., Ghia, P., Giustina, A., Lazzarin, A., Nuti, R., Pardi, R., & Stella, A. P. (2021). *Medicina Interna Sistematica* (Ottava Edizione) ISBN: 8821450953.
11. Summers C, Todd RS, Vercruysse GA, Moore FA. Acute Respiratory Failure. *Perioperative Medicine*. 2022;576-586. https://doi.org/10.1016/B978-0-323-56724-4.00039-3 .

12. Fernandes, C. J., Luppino Assad, A. P., Alves-Jr, J. L., Jardim, C., & De Souza, R. (2019). Pulmonary Embolism and Gas Exchange. *Respiration*, *98*(3), 253–262. https://doi.org/10.1159/000501342
13. Drechsler, M., & Morris, J. (2026). *Carbon Dioxide Narcosis*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK551620/
14. Macintyre, N. R. (2023). Acute Hypercapnic Respiratory Failure in COPD. *Respiratory Care*, *68*(7), 973–982. https://doi.org/10.4187/respcare.10560
15. McMullen, S. M., & Patrick, W. (2013). Cyanosis. In *American Journal of Medicine* (Vol. 126, Number 3, pp. 210–212). Elsevier Inc. https://doi.org/10.1016/j.amjmed.2012.11.004
16. Castro, D., Patil, S. M., Zubair, M., & Keenaghan, M. (2024). *Arterial Blood Gas*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK536919/
17. Breen, P. H. (2001). Arterial blood gas and pH analysis: Clinical approach and interpretation. *Anesthesiology Clinics of North America*, *19*(4), 885–906. https://doi.org/10.1016/S0889-8537(01)80014-6
18. DeVos, E., & Jacobson, L. (2016). Approach to Adult Patients with Acute Dyspnea. In *Emergency Medicine Clinics of North America* (Vol. 34, Number 1, pp. 129–149). W.B. Saunders. https://doi.org/10.1016/j.emc.2015.08.008
19. Santus, P., Radovanovic, D., Saad, M., Zilianti, C., Coppola, S., Chiumello, D. A., & Pecchiari, M. (2023). Acute dyspnea in the emergency department: a clinical review. In *Internal and Emergency Medicine* (Vol. 18, Number 5, pp. 1491–1507). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s11739-023-03322-8
20. Islam, M., Levitus, M., Eisen, L., Shiloh, A. L., & Fein, D. (2020). Lung Ultrasound for the Diagnosis and Management of Acute Respiratory Failure. In *Lung* (Vol. 198, Number 1). Springer. https://doi.org/10.1007/s00408-019-00309-1
21. Lichtenstein, D. A., & Mezière, G. A. (2008). Relevance of lung ultrasound in the diagnosis of acute respiratory failure the BLUE protocol. *Chest*, *134*(1), 117–125. https://doi.org/10.1378/chest.07-2800
22. Wallbridge, P., Steinfort, D., Tay, T. R., Irving, L., & Hew, M. (2018). Diagnostic chest ultrasound for acute respiratory failure. In *Respiratory Medicine* (Vol. 141, pp. 26–36). W.B. Saunders Ltd. https://doi.org/10.1016/j.rmed.2018.06.018
23. Cosentini, R., Aliberti, S., & Brambilla, A. M. (2010). *L'ABC della ventilazione meccanica non invasiva in urgenza* (Seconda Edizione). ISBN:9788838639487

24. Bello, G., Ionescu Maddalena, A., Giammatteo, V., & Antonelli, M. (2018). Noninvasive Options. In *Critical Care Clinics* (Vol. 34, Number 3, pp. 395–412). W.B. Saunders. https://doi.org/10.1016/j.ccc.2018.03.007
25. Walter, J. M., Corbridge, T. C., & Singer, B. D. (2018). Invasive Mechanical Ventilation. *Southern Medical Journal*, *111*(12), 746–753. https://doi.org/10.14423/SMJ.0000000000000905
26. Spinazzola, G., Ferrone, G., Michi, T., Torrini, F., Postorino, S., Sbaraglia, F., Gulmini, L., Antonelli, M., Conti, G., & Spadaro, S. (2024). Focus on the Role of Non-Invasive Respiratory Support (NRS) during Palliative Care in Patients with Life-Limiting Respiratory Disease. In *Journal of Clinical Medicine* (Vol. 13, Number 17). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/jcm13175165
27. Pearce, A. K., Nemati, S., Goligher, E. C., Hough, C. L., Holder, A. L., Wardi, G., Yang, P., Boussina, A., Lyons, P. G., Sahetya, S., Malhotra, A., & Rogers, A. (2025). Can we predict the future of respiratory failure prediction? *Critical Care*, *29*(1). https://doi.org/10.1186/s13054-025-05484-7
28. Narendra, D. K., Hess, D. R., Sessler, C. N., Belete, H. M., Guntupalli, K. K., Khusid, F., Carpati, C. M., Astiz, M. E., & Raoof, S. (2017). Update in Management of Severe Hypoxemic Respiratory Failure. In *Chest* (Vol. 152, Number 4, pp. 867–879). Elsevier Inc. https://doi.org/10.1016/j.chest.2017.06.039
29. Roca, O., Caralt, B., Messika, J., Samper, M., Sztrymf, B., Hernández, G., García-De-Acilu, M., Frat, J. P., Masclans, J. R., & Ricard, J. D. (2019). An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy. *American Journal of Respiratory and Critical Care Medicine*, *199*(11), 1368–1376. https://doi.org/10.1164/rccm.201803-0589OC
30. Frat, J. P., Ragot, S., Coudroy, R., Constantin, J. M., Girault, C., Prat, G., Boulain, T., Demoule, A., Ricard, J. D., Razazi, K., Lascarrou, J. B., Devaquet, J., Mira, J. P., Argaud, L., Chakarian, J. C., Fartoukh, M., Nseir, S., Mercat, A., Brochard, L., … Thille, A. W. (2018). Predictors of intubation in patients with acute hypoxemic respiratory failure treated with a noninvasive oxygenation strategy. *Critical Care Medicine*, *46*(2), 208–215. https://doi.org/10.1097/CCM.0000000000002818
31. Goh, K. J., Chai, H. Z., Ong, T. H., Sewa, D. W., Phua, G. C., & Tan, Q. L. (2020). Early prediction of high flow nasal cannula therapy outcomes using a modified ROX

index incorporating heart rate. *Journal of Intensive Care*, *8*(1). https://doi.org/10.1186/s40560-020-00458-z

32. Roca, O., Messika, J., Caralt, B., García-de-Acilu, M., Sztrymf, B., Ricard, J. D., & Masclans, J. R. (2016). Predicting success of high-flow nasal cannula in pneumonia patients with hypoxemic respiratory failure: The utility of the ROX index. *Journal of Critical Care*, *35*, 200–205. https://doi.org/10.1016/j.jcrc.2016.05.022
33. Duan, J., Han, X., Bai, L., Zhou, L., & Huang, S. (2017). Assessment of heart rate, acidosis, consciousness, oxygenation, and respiratory rate to predict noninvasive ventilation failure in hypoxemic patients. *Intensive Care Medicine*, *43*(2), 192–199. https://doi.org/10.1007/s00134-016-4601-3
34. Duan, J., Chen, L., Liu, X., Bozbay, S., Liu, Y., Wang, K., Esquinas, A. M., Shu, W., Yang, F., He, D., Chen, Q., Wei, B., Chen, B., Li, L., Tang, M., Yuan, G., Ding, F., Huang, T., Zhang, Z., … Mina, B. (2022). An updated HACOR score for predicting the failure of noninvasive ventilation: a multicenter prospective observational study. *Critical Care*, *26*(1). https://doi.org/10.1186/s13054-022-04060-7
35. Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. B. (2019). Machine learning: applications of artificial intelligence to imaging and diagnosis. In *Biophysical Reviews* (Vol. 11, Number 1, pp. 111–118). Springer Verlag. https://doi.org/10.1007/s12551-018-0449-9
36. Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Peter Campbell, J. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, *9*(2). https://doi.org/10.1167/tvst.9.2.14
37. Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. In *Diagnostics* (Vol. 13, Number 15). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics13152582
38. Fodor, J. A., & Pylyshyn, Z. W. (1988). *Connectionism and cognitive architecture: a critical analysis*. https://doi.org/10.1016/0010-0277(88)90031-5
39. Quintarelli, S. (2020). *Intelligenza Artificiale. Cos'è davvero, come funziona, che effetti avrà*. Bollati Boringhieri. ISBN: 9788833935423

40. Nielsen, M. (2019). *Neural Networks and Deep Learning*. http://neuralnetworksanddeeplearning.com/chap1.html
41. Sadiku, M. N. O., Ajayi, S. A., & Sadiku, J. O. (2025). Machine Learning: An Overview the Creative Commons Attribution License (CC BY 4.0). In *International Journal of Trend in Scientific Research and Development*. www.ijtsrd.com/papers/ijtsrd97693.pdf
42. *What Is Machine Learning in Health Care?* (2025). Coursera. https://www.coursera.org/articles/machine-learning-in-health-care
43. Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386. https://doi.org/10.21275/art20203995
44. Kavlakoglu, E. (n.d.). *What is a decision tree?* International Business Machines Corporation. https://www.ibm.com/think/topics/decision-trees
45. What is overfitting? (n.d.). *International Business Machines Corporation*. https://www.ibm.com/think/topics/overfitting
46. Salian, I. (2018). *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?* NVIDIA Blog. https://www.edge-ai-vision.com/2018/08/supervize-me-whats-the-difference-between-supervised-unsupervised-semi-supervised-and-reinforcement-learning/
47. Mateussi, N., Rogers, M. P., Grimsley, E. A., Read, M., Parikh, R., Pietrobon, R., & Kuo, P. C. (2024). Clinical Applications of Machine Learning. *Annals of Surgery Open*, *5*(2), e423. https://doi.org/10.1097/as9.0000000000000423
48. Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, *17*(11). https://doi.org/10.1111/cts.70056
49. What is XGBoost? (n.d.). *International Business Machines Corporation*. https://www.ibm.com/think/topics/xgboost
50. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. Bin, Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, *637*(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6
51. Chen, W., Tian, Y., Liao, N., Deng, Y., Jiang, D., & Cao, D. (2026). TabPFN Opens New Avenues for Small-Data Tabular Learning in Drug Discovery. *Journal of*

Chemical Information and Modeling, *66*(7), 3525–3539. https://doi.org/10.1021/acs.jcim.5c02823

52. Wu, X., Ritter, A., & Xu, W. (2025). *Tabular Data Understanding with LLMs: A Survey of Recent Advances and Challenges*. http://arxiv.org/abs/2508.00217
53. Jiang, J.-P., Liu, S.-Y., Cai, H.-R., Zhou, Q., & Ye, H.-J. (2025). *Representation Learning for Tabular Data: A Comprehensive Survey*. http://arxiv.org/abs/2504.16109
54. Jang, E., & Lee, K. Y. (2024). GBGNN: Gradient Boosted Graph Neural Networks. *Journal of Information Processing Systems*, *20*(4), 501–513. https://doi.org/10.3745/JIPS.04.0315
55. Shen, J., Li, L., Dery, L. M., Staten, C., Khodak, M., Neubig, G., & Talwalkar, A. (n.d.). *Cross-Modal Fine-Tuning: Align then Refine*. https://arxiv.org/abs/2302.05738
56. Bendavid, I., Statlender, L., Shvartser, L., Teppler, S., Azullay, R., Sapir, R., & Singer, P. (2022). A novel machine learning model to predict respiratory failure and invasive mechanical ventilation in critically ill patients suffering from COVID-19. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-14758-x
57. Liu, S.-Y., & Ye, H.-J. (2025). *TabPFN Unleashed: A Scalable and Effective Solution to Tabular Classification Problems* (Vol. 267). https://arxiv.org/abs/2502.02527
58. McCarter, C. (2025). *What exactly has TabPFN learned to do?* http://arxiv.org/abs/2502.08978
59. Zhang, Q., Tan, Y. S., Tian, Q., & Li, P. (2025). *TabPFN: One Model to Rule Them All?* http://arxiv.org/abs/2505.20003
60. Güler, R. A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., & Kokkinos, I. (n.d.). *DenseReg: Fully Convolutional Dense Shape Regression In-the-Wild*. Retrieved http://alpguler.com/DenseReg.html
61. Yu, H., Saffaran, S., Tonelli, R., Laffey, J. G., Zhang, Q., Esquinas, A. M., de Lima, L. M., Kawano-Dourado, L., Maia, I. S., Cavalcanti, A. B., Clini, E., & Bates, D. G. (2025). 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. *Critical Care*, *29*(1). https://doi.org/10.1186/s13054-025-05765-1
62. Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., & Gandomi, A. H. (2022). Machine learning in medical

applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, *145*. https://doi.org/10.1016/j.compbiomed.2022.105458

63. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, *24*(11), 1716–1720. https://doi.org/10.1038/s41591-018-0213-5
64. Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., Bock, C., Horn, M., Moor, M., Rieck, B., Zimmermann, M., Bodenham, D., Borgwardt, K., Rätsch, G., & Merz, T. M. (2020). Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, *26*(3), 364–373. https://doi.org/10.1038/s41591-020-0789-4
65. Akgün, K. B., Al-Otaibi, H. M., Chiappa, G. R., Alqarni, A. A., Troxell, D., Przybylski, F., & Esquinas, A. M. (2026). Machine learning-based prediction models for noninvasive respiratory support failure in acute respiratory failure: a systematic review and meta-analysis. *Frontiers in Medicine*, *13*. https://doi.org/10.3389/fmed.2026.1775670
66. Notaro, S., Giurazza, R., Imparato, A., Falso, F., Virno, S., Iovino, T. P., Caiazzo, M., Golino, L., Russo, G., Di Costanzo, E., Barberio, M., Sorrentino, M., Fiorentino, G., La Cerra, G., Esposito, C., Corcione, A., Esquinas, A. M., & Piscitelli, E. (2026). Predicting failure of extubation and non-invasive respiratory support in critically ill patients: clinical complexity, limitations of traditional indices, and machine learning perspectives. In *Frontiers in Medicine* (Vol. 13). Frontiers Media SA. https://doi.org/10.3389/fmed.2026.1791505
67. Wang, Z., Chao, Y., Xu, M., Zhao, W., & Hu, X. (2024). Machine learning prediction of the failure of high-flow nasal oxygen therapy in patients with acute respiratory failure. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-52061-z
68. Temte, B., Carey, K. A., Spicer, A., Shah, N., Winslow, C., Gilbert, E., Afshar, M., Edelson, D., & Churpek, M. (2025). Development and external validation of a machine learning model to predict high flow nasal cannula failure. *BMJ Open Respiratory Research*, *12*(1). https://doi.org/10.1136/bmjresp-2025-003355
69. Yu, H., Saffaran, S., Tonelli, R., Laffey, J. G., Esquinas, A. M., de Lima, L. M., Kawano-Dourado, L., Maia, I. S., Cavalcanti, A. B., Clini, E., & Bates, D. G. (2025). Machine learning models compared with current clinical indices to predict the

outcome of high flow nasal cannula therapy in acute hypoxemic respiratory failure. *Critical Care*, *29*(1). https://doi.org/10.1186/s13054-025-05336-4

70. Essay, P. T., Mosier, J. M., Nayebi, A., Fisher, J. M., & Subbian, V. (2023). Predicting Failure of Noninvasive Respiratory Support Using Deep Recurrent Learning. *Respiratory Care*, *68*(4), 488–496. https://doi.org/10.4187/respcare.10382
71. Rockwood, K., Song, X., MacKnight, C., Bergman, H., Hogan, D. B., McDowell, I., & Mitnitski, A. (2005). A global clinical measure of fitness and frailty in elderly people. *CMAJ. Canadian Medical Association Journal*, *173*(5), 489–495. https://doi.org/10.1503/cmaj.050051
72. Clegg, A. (2011). CME Geriatric medicine. In *Clinical Medicine* (Vol. 11, Number 1). https://doi.org/10.7861/clinmedicine.11-1-72
73. Ferrante, L. E., & Szczeklik, W. (2024). Frailty is crucial in FORECASTing outcomes in critical care. In *Intensive Care Medicine* (Vol. 50, Number 7, pp. 1119–1122). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00134-024-07518-0
74. Melo, R. C. de, Cervato, C. J., Aprahamian, I., Gavin, J., Robinson, K., Frost, R., Azevedo, P. S., Villas-Boas, P., Hinsliff-Smith, K., & Gordon, A. L. (2020). Identifying and managing frailty in Brazil: a scoping review protocol. *JBI Evidence Synthesis*, *18*(4), 849–856. https://doi.org/10.11124/JBISRIR-D-19-00188
75. Cohen, C. I., Benyaminov, R., Rahman, M., Ngu, D., & Reinhardt, M. (2023). Frailty: A Multidimensional Biopsychosocial Syndrome. In *Medical Clinics of North America* (Vol. 107, Number 1, pp. 183–197). W.B. Saunders. https://doi.org/10.1016/j.mcna.2022.04.006
76. Farhat, J. S., Velanovich, V., Falvo, A. J., Horst, H. M., Swartz, A., Patton, J. H., & Rubinfeld, I. S. (2012). Are the frail destined to fail? Frailty index as predictor of surgical morbidity and mortality in the elderly. *Journal of Trauma and Acute Care Surgery*, *72*(6), 1526–1531. https://doi.org/10.1097/TA.0b013e3182542fab
77. Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W. J., Burke, G., & Mcburnie, M. A. (2001). Frailty in Older Adults: Evidence for a Phenotype. In *Journal of Gerontology: MEDICAL SCIENCES Copyright* (Vol. 56, Number 3). https://academic.oup.com/biomedgerontology/article-abstract/56/3/M146/545770

78. Mitnitski, A. B., Mogilner, A. J., & Rockwood, K. (2001). Accumulation of deficits as a proxy measure of aging. *TheScientificWorldJournal*, *1*, 323–336. https://doi.org/10.1100/tsw.2001.58
79. Morley, J. E., Malmstrom, T. K., & Miller, D. K. (n.d.). *A simple frailty questionnaire (frail) predicts outcomes in middle aged african americans*. https://doi.org/10.1007/s12603-012-0084-2
80. Petty, D. R., House, A., Knapp, P., Raynor, T., & Zermansky, A. (2006). Prevalence, duration and indications for prescribing of antidepressants in primary care [3]. In *Age and Ageing* (Vol. 35, Number 5, pp. 523–526). https://doi.org/10.1093/ageing/afl023
81. Perna, S., Francis, M. D. A., Bologna, C., Moncaglieri, F., Riva, A., Morazzoni, P., Allegrini, P., Isu, A., Vigo, B., Guerriero, F., & Rondanelli, M. (2017). Performance of Edmonton Frail Scale on frailty assessment: its association with multi-dimensional geriatric conditions assessed with specific screening tools. *BMC Geriatrics*, *17*(1), 1–8. https://doi.org/10.1186/s12877-016-0382-3
82. Schuurmans, H., Steverink, N., Lindenberg, S., Frieswijk, N., & Slaets, J. P. J. (2004). *Rapid Communication Old or Frail: What Tells Us More?* http://biomedgerontology.oxfordjournals.org/
83. Marcucci, M., Franchi, C., Nobili, A., Mannucci, P. M., Ardoino, I., Tettamanti, M., Pasina, L., Perticone, F., Salerno, F., Corrao, S., Marengoni, A., Licata, G., Violi, F., Corazza, G. R., Eldin, T. K., Di Blanca, M. P. D., Djade, C. D., Cortesi, L., Prisco, D., … Monica, R. P. (2017). Defining aging phenotypes and related outcomes: Clues to recognize frailty in hospitalized older patients. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *72*(3), 395–402. https://doi.org/10.1093/gerona/glw188
84. Zhou, Y., & Ma, L. (2022). Intrinsic Capacity in Older Adults: Recent Advances. In *Aging and Disease* (Vol. 13, Number 2, pp. 353–359). International Society on Aging and Disease. https://doi.org/10.14336/AD.2021.0818
85. Bandopadhyay, R., Vaidya, N., Ayaz, A., Maghamifar, J., Momtaz, A., Huda, M., & Ansari, Z. (2025). Correlation of Clinical Frailty Scale and Outcomes in Patients Treated With Non-invasive Ventilation: A Retrospective Cohort Study. *Cureus*. https://doi.org/10.7759/cureus.95641
86. Mort, M., Raju, A., & Mohan, N. (2025). Frailty as a Predictor of Outcomes in Optiflow. *Cureus*. https://doi.org/10.7759/cureus.95828

87. Fernando, S. M., McIsaac, D. I., Rochwerg, B., Bagshaw, S. M., Muscedere, J., Munshi, L., Ferguson, N. D., Seely, A. J. E., Cook, D. J., Dave, C., Tanuseputro, P., & Kyeremanteng, K. (2019). Frailty and invasive mechanical ventilation: association with outcomes, extubation failure, and tracheostomy. *Intensive Care Medicine*, *45*(12), 1742–1752. https://doi.org/10.1007/s00134-019-05795-8
88. Saxena, S., Gupta, P., Panwar, P., Jain, A., Jain, S. S., Jain, R., Gupta, D., Meena, M., Acharya, H., & Jain, R. (2025). Impact of the Clinical Frailty Score on Outcomes of Critically Ill Patients in a Tertiary Care ICU. *Indian Journal of Critical Care Medicine*, *29*(4), 320–326. https://doi.org/10.5005/jp-journals-10071-24949
89. Stefan, M., Asghar, A., Shieh, M.-S., Demir-Yavuz, S., & Steingrub, J. S. (2022). The Association of Frailty With Long-Term Outcomes in Patients With Acute Respiratory Failure Treated With Noninvasive Ventilation. *Cureus*. https://doi.org/10.7759/cureus.33143
90. Chen, Y., Zhang, X., Chen, Y., & Tong, Z. (2024). Assessment of bidirectional relationships between frailty and acute respiratory distress syndrome: a bidirectional Mendelian Randomization study. *BMC Geriatrics*, *24*(1). https://doi.org/10.1186/s12877-024-05579-w
91. Kozesinski-Nakatani, A. C., Bernardelli, R. S., Vieira, J. C. F., Martins-Junior, M. J., Tiba, M. N., Barros, V. S., Tiba, T., Konzen, B. L., Janiscki, K., Réa-Neto, Á., & de Lima Moser, A. D. (2025). Evaluation of the clinical frailty scale for predicting mortality or functional dependence at ICU discharge: A cohort study. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-23243-0
92. Blythe, R., White, N. M., Brown, C., Hillman, K., & Barnett, A. G. (2025). Validation of the Clinical Frailty Scale for predicting 90-day mortality in hospitalised older adults screened as at risk of nearing the end of life in Queensland, Australia: a multisite observational study. *BMJ Open* , *15*(11). https://doi.org/10.1136/bmjopen-2025-108419
93. Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation Studies in Artificial Neural Networks*. http://arxiv.org/abs/1901.08644
94. Tonelli, R., Fantini, R., Tabbì, L., Castaniere, I., Pisani, L., Pellegrino, M. R., Della Casa, G., D'Amico, R., Girardis, M., Nava, S., Clini, E. M., & Marchioni, A. (2020). Early Inspiratory Effort Assessment by Esophageal Manometry Predicts Noninvasive Ventilation Outcome in *De Novo* Respiratory Failure. A Pilot

Study. *American journal of respiratory and critical care medicine*, *202*(4), 558–567. https://doi.org/10.1164/rccm.201912-2512OC

95. Tonelli, R., Busani, S., Tabbì, L., Fantini, R., Castaniere, I., Biagioni, E., Mussini, C., Girardis, M., Clini, E., & Marchioni, A. (2021). Inspiratory Effort and Lung Mechanics in Spontaneously Breathing Patients with Acute Respiratory Failure due to COVID-19: A Matched Control Study. *American journal of respiratory and critical care medicine*, *204*(6), 725–728. https://doi.org/10.1164/rccm.202104-1029LE
96. Tonelli, R., Cortegiani, A., Fantini, R., Tabbì, L., Castaniere, I., Bruzzi, G., Busani, S., Ball, L., Clini, E., & Marchioni, A. (2023). Accuracy of Nasal Pressure Swing to Predict Failure of High-Flow Nasal Oxygen in Patients with Acute Hypoxemic Respiratory Failure. *American journal of respiratory and critical care medicine*, *207*(6), 787–789. https://doi.org/10.1164/rccm.202210-1848LE
97. Tonelli, R., Bruzzi, G., Fantini, R., Tabbì, L., Moretti, A., Livrieri, F., Castaniere, I., Cerri, S., Puggioni, D., Guidotti, F., Bellesia, G., Marchioni, A., & Clini, E. (2025). Assessment of nasal pressure swing predicts respiratory support dependency in patients with hypoxic respiratory failure. *European journal of internal medicine*, *135*, 91–97. https://doi.org/10.1016/j.ejim.2025.02.016
98. Tabbì, L., Marchioni, A., Messina, J., Bombelli, S., Puggioni, D., Fantini, R., Bruzzi, G., Cerri, S., Beghè, B., Clini, E., Tonelli, R., & MUSE-COPD study group (2026). Ultrasound-based evaluation of inspiratory muscle effort in patients with type 2 respiratory failure secondary to exacerbation of chronic obstructive pulmonary disease. *European journal of internal medicine*, *145*, 106649. https://doi.org/10.1016/j.ejim.2025.106649
99. Yu, H., Saffaran, S., Maia, I. S., Clini, E., Bates, D. G., & NIVPredict study group (2025). Early prediction of non-invasive ventilation outcome using the TabPFN machine learning model: a multi-centre validation study. *Intensive care medicine*, *51*(8), 1542–1544. https://doi.org/10.1007/s00134-025-08025-6
100. Yu, H., Saffaran, S., Ali, A., Henry, C., Mustfa, N., Thomas, A., Rajhan, A., Isrhad, S., Weaver, L., Tonelli, R., Menga, L. S., Zhang, Q., Samadi, M. E., Schuppert, A., Laffey, J. G., Camporota, L., Esquinas, A. M., Grieco, D. L., Antonelli, M., de Lima, L. M., … Bates, D. G. (2026). In-hospital testing of NIVPredict - an AI tool for early prediction of non-invasive ventilation outcome in

acute respiratory failure. *Critical care (London, England)*, *30*(1), 124. https://doi.org/10.1186/s13054-026-05894-1

101. Parra-Santander, D., & Brusilovsky, P. (2010). *Improving Collaborative Filtering in Social Tagging Systems for the Recommendation of Scientific Articles*. https://doi.org/10.1007/978-3-642-25274-7_47

102. https://shap.readthedocs.io/en/latest/generated/shap.KernelExplainer.html

103. Kavlakoglu, E., & Winland, V. (n.d.). *Cos'è il k-means clustering?* International Business Machines Corporation. https://www.ibm.com/it-it/think/topics/k-means-clustering

104. Assel, M., Sjoberg, D. D., & Vickers, A. J. (2017). The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnostic and Prognostic Research*, *1*(1). https://doi.org/10.1186/s41512-017-0020-3

105. Montoneri, G., Noto, P., Trovato, F. M., Mangano, G., Malatino, L., & Carpinteri, G. (2019). Outcomes of non-invasive ventilation in "very old" patients with acute respiratory failure: A retrospective study. *Emergency Medicine Journal*, *36*(5), 303–305. https://doi.org/10.1136/emermed-2018-207563

106. Schortgen, F., Follin, A., Piccari, L., Roche-Campo, F., Carteaux, G., Taillandier-Heriche, E., Krypciak, S., Thille, A. W., Paillaud, E., & Brochard, L. (2012). Results of noninvasive ventilation in very old patients. *Annals of Intensive Care*, *2*(1), 1–9. https://doi.org/10.1186/2110-5820-2-5

107. Bourke, S. C., Piraino, T., Pisani, L., Brochard, L., & Elliott, M. W. (2018). Beyond the guidelines for non-invasive ventilation in acute respiratory failure: implications for practice. In *The Lancet Respiratory Medicine* (Vol. 6, Number 12, pp. 935–947). Lancet Publishing Group. https://doi.org/10.1016/S2213-2600(18)30388-6

108. Rochwerg, B., Brochard, L., Elliott, M. W., Hess, D., Hill, N. S., Nava, S., Navalesi, P., Antonelli, M., Brozek, J., Conti, G., Ferrer, M., Guntupalli, K., Jaber, S., Keenan, S., Mancebo, J., Mehta, S., & Raoof, S. (2017). Official ERS/ATS clinical practice guidelines: Noninvasive ventilation for acute respiratory failure. In *European Respiratory Journal* (Vol. 50, Number 4). European Respiratory Society. https://doi.org/10.1183/13993003.02426-2016

109. Abulesz, Y. T., & Haugk, M. (2023). Impact of noninvasive ventilation at a municipal emergency department on ICU admissions. *Wiener Klinische Wochenschrift*, *135*(1–2), 33. https://doi.org/10.1007/s00508-022-02111-1

110. Grieco, D. L., Menga, L. S., Raggi, V., Bongiovanni, F., Anzellotti, G. M., Tanzarella, E. S., Bocci, M. G., Mercurio, G., Dell'Anna, A. M., Eleuteri, D., Bello, G., Maviglia, R., Conti, G., Maggiore, S. M., & Antonelli, M. (2020). Physiological comparison of high-flow nasal cannula and helmet noninvasive ventilation in acute hypoxemic respiratory failure. *American Journal of Respiratory and Critical Care Medicine*, *201*(3), 303–312. https://doi.org/10.1164/rccm.201904-0841OC

111. Qin, J., Wang, G., Liao, Y., Shang, W., & Han, D. (2025). High flow nasal therapy versus noninvasive ventilation for AECOPD with acute hypercapnic respiratory failure: a meta-analysis of randomized controlled trials. In *Annals of Intensive Care* (Vol. 15, Number 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s13613-025-01480-w

112. Frat, J. P., Coudroy, R., Marjanovic, N., & Thille, A. W. (2017). High-flow nasal oxygen therapy and noninvasive ventilation in the management of acute hypoxemic respiratory failure. In *Annals of Translational Medicine* (Vol. 5, Number 14). AME Publishing Company. https://doi.org/10.21037/atm.2017.06.52

113. Vargas, F., Saint-Leger, M., Boyer, A., Bui, N. H., & Hilbert, G. (2015). Physiologic effects of high-flow nasal Cannula oxygen in critical care subjects. *Respiratory Care*, *60*(10), 1369–1376. https://doi.org/10.4187/respcare.03814

114. Menga, L. S., Balzani, E., Gelormini, C., Mastropietro, C., Volpe, N., Del Signore, R., Mondello, M., Cardu, A., Delle Cese, L., Rosà, T., D'Arrigo, S., Maggiore, S. M., Brochard, L., Antonelli, M., Bellani, G., & Grieco, D. L. (2025). Physiological effects of noninvasive respiratory support strategies in adults with acute hypoxemic respiratory failure: a systematic review and network meta-analysis. *Critical Care*, *29*(1). https://doi.org/10.1186/s13054-025-05670-7

115. Vaz Fragoso, C. A., Enright, P. L., McAvay, G., Van Ness, P. H., & Gill, T. M. (2012). Frailty and respiratory impairment in older persons. *American Journal of Medicine*, *125*(1), 79–86. https://doi.org/10.1016/j.amjmed.2011.06.024

116. https://frailtyscience.org/pulmonary-function-and-frailty/

117. Vidal, M. B., Pegorari, M. S., Santos, E. C., Matos, A. P., Pinto, A. C. P. N., & Ohara, D. G. (2020). Respiratory muscle strength for discriminating frailty in community-dwelling elderly: a cross-sectional study. *Archives of Gerontology and Geriatrics*, *89*. https://doi.org/10.1016/j.archger.2020.104082

118. Nagano, A., Wakabayashi, H., Maeda, K., Kokura, Y., Miyazaki, S., Mori, T., & Fujiwara, D. (2021). Respiratory Sarcopenia and Sarcopenic Respiratory

Disability: Concepts, Diagnosis, and Treatment. *Journal of Nutrition, Health and Aging*, *25*(4), 507–515. https://doi.org/10.1007/s12603-021-1587-5

119. Gallifant, J., Zhang, J., Del Pilar Arias Lopez, M., Zhu, T., Camporota, L., Celi, L. A., & Formenti, F. (2022). Artificial intelligence for mechanical ventilation: systematic review of design, reporting standards, and bias. *British journal of anaesthesia*, *128*(2), 344. https://doi.org/10.1016/j.bja.2021.09.025

120. Price, D., Brusselle, G., Roche, N., Freeman, D., & Chisholm, A. (2015). Real-world research and its importance in respiratory medicine. *Breathe*, *11*(1), 33. https://doi.org/10.1183/20734735.015414

121. Ong, S. W. X., Tong, S. Y. C., & Daneman, N. (2023). Are we enrolling the right patients? A scoping review of external validity and generalizability of clinical trials in bloodstream infections. In *Clinical Microbiology and Infection* (Vol. 29, Number 11, pp. 1395–1399). Elsevier B.V. https://doi.org/10.1016/j.cmi.2023.08.019

122. Patra, S. (2026). *Corrected Resampled t-test*. https://suchibrata.in/blog/statistical-tests/corrected-resampled-t-test

123. Kiani, J. (2019). *Using the Corrected Paired Student's t-test for comparing Machine Learning Models*. https://medium.com/analytics-vidhya/using-the-corrected-paired-students-t-test-for-comparing-the-performance-of-machine-learning-dc6529eaa97f

124. Schisterman, E. F., Faraggi, D., Reiser, B., & Hu, J. (2008). Youden index and the optimal threshold for markers with mass at zero. Statistics in Medicine, 27(2), 297–315. https://doi.org/10.1002/sim.2993

125. Bury, T. (2024). Interpretable ML 3: What nobody tells you about SHAP values. https://bury-thomas.medium.com/interpretability-ml-3-what-nobody-tells-you-about-shap-values-8295f0db3f9d

126. Ng, W. Y., Wang, L. R., Liu, S., & Fan, X. (2025). *Causal SHAP: Feature Attribution with Dependency Awareness through Causal Discovery*. http://arxiv.org/abs/2509.00846

7 Acknowledgments

Desidero innanzitutto ringraziare il mio relatore, il Prof. Roberto Tonelli, per avermi guidato, stimolato e incoraggiato nel mio percorso di crescita professionale nell'ambiente della Pneumologia. Ringrazio inoltre il Prof. Enrico Clini per avermi permesso di entrare nella Struttura Complessa di Malattie dell'Apparato Respiratorio. Ringrazio infine tutto il reparto per avermi accolto in questi anni.