

POLITECNICO DI TORINO

Facoltà di Ingegneria

Corso di Laurea Magistrale in Ingegneria Biomedica

Tesi di Laurea Magistrale

**Data science e machine learning per  
l'individuazione del tempo ottimale per  
svezzare un paziente dal ventilatore meccanico**



**Relatore:**

Prof. Gabriella Balestra

**Co-Relatore:**

Prof. Samanta Rosati

**Candidato:**

Sara Biscione

Dicembre 2020



## Sommario

La ventilazione meccanica è la principale strategia terapeutica impiegata per fornire supporto alla funzione respiratoria dei pazienti incapaci di respirare autonomamente. Il numero dei pazienti ricoverati in terapia intensiva (ICU) che necessitano di ventilazione meccanica invasiva ha avuto un andamento crescente negli ultimi anni, a causa della sempre più elevata vita media della popolazione, e ha raggiunto un picco nell'anno 2020 in seguito alla diffusione del nuovo Coronavirus (SARS-CoV-2).

Nella cura dei pazienti intubati in condizioni critiche lo svezzamento dal ventilatore meccanico è un elemento essenziale e universale. È cruciale, infatti, determinare il momento appropriato per svezzare un paziente dalla ventilazione meccanica invasiva. Ogni nuovo giorno di ventilazione, senza che avvenga un'estubazione con successo, è associato a una mortalità progressivamente crescente. Inoltre, una ventilazione prolungata aumenta la probabilità di contrarre infezioni polmonari associate al ventilatore, i costi ospedalieri e le risorse dell'unità di terapia intensiva utilizzate. Emerge, dunque, l'idea di sviluppare un sistema CAD (computer-aided diagnosis) per supportare il medico nell'individuazione del momento ottimale per svezzare il paziente dal ventilatore meccanico.

Il progetto di tesi si pone, quindi, l'obiettivo di indagare i modelli di classificazione adatti al problema clinico in questione, esplorando gli esistenti studi. A questo scopo, è necessario valutare le variabili che descrivono al meglio le condizioni cliniche del paziente da svezzare, usufruendo del database dedicato alla medicina intensiva MIMIC-III (Medical Information Mart for Intensive Care). La prima fase della realizzazione del lavoro di tesi è stata la costruzione del dataset e il conseguente data cleaning, passando per la selezione dei soggetti, la determinazione della classe a cui appartengono (estubazione fallita / estubazione riuscita) arrivando all'estrazione delle variabili caratterizzanti ogni evento di ventilazione dei pazienti. In una fase successiva, è stata effettuata una suddivisione del dataset creando, in particolare, un training set bilanciato e rappresentativo dell'intero dataset. Questo è stato utile alla costruzione di due tipologie di classificatore, le Reti Neurali e il Light Gradient Boosting Machine. Per la prima tipologia di modello è stata effettuata una fase di allenamento analizzando diverse architetture, un confronto e valutazione di queste e infine, una feature selection combinata all'ottimizzazione della struttura della rete neurale, tramite un algoritmo genetico. Affrontando la costruzione del LightGBM, invece, è stato effettuato un allenamento del modello a seguito del tuning degli iperparametri, una valutazione e interpretazione dei risultati del modello tramite importanza e shap value e infine, una forward feature selection.

## Indice

<b>1</b>	<b>Introduzione</b> .....	3
	1.1 Scopo del progetto di tesi .....	4
	1.2 Situazione attuale .....	4
	1.3 Passaggi seguiti nel progetto di tesi .....	7
<b>2</b>	<b>Database MIMIC-III</b> .....	8
<b>3</b>	<b>Costruzione del dataset</b> .....	11
	3.1 Selezione dei soggetti.....	11
	3.1.1 Estrazione degli eventi di ventilazione meccanica invasiva .....	11
	3.1.2 Scelta e applicazione dei criteri di inclusione ed esclusione.....	12
	3.2 Determinazione della classe di ogni evento di ventilazione.....	14
	3.3 Calcolo ed estrazione delle variabili .....	15
<b>4</b>	<b>Data cleaning</b> .....	18
<b>5</b>	<b>Suddivisione del dataset</b> .....	21
<b>6</b>	<b>Costruzione del classificatore Light Gradient Boosting Machine</b> .....	25
	6.1 Apprendimento e validazione .....	30
	6.2 Interpretazione del modello .....	34
	6.3 Forward feature selection .....	36
<b>7</b>	<b>Costruzione del classificatore Rete Neurale</b> .....	38
	7.1 Apprendimento e validazione .....	38
	7.2 Feature selection: algoritmo genetico .....	41
<b>8</b>	<b>Discussione e conclusioni</b> .....	47
	<b>Appendice A</b> .....	53
	<b>Appendice B</b> .....	57
	<b>Bibliografia</b> .....	64

## Capitolo 1

### Introduzione

La ventilazione meccanica è la principale strategia terapeutica impiegata per fornire supporto alla funzione respiratoria dei pazienti incapaci di respirare autonomamente. Il numero dei pazienti ricoverati in terapia intensiva (ICU) che necessitano di ventilazione meccanica ha avuto un andamento crescente negli ultimi anni, a causa della sempre più elevata vita media della popolazione [1], e ha raggiunto un picco nell'anno 2020 in seguito alla diffusione del nuovo Coronavirus (SARS-CoV-2).

Il supporto ventilatorio può essere fornito al paziente con modalità non invasiva, tramite l'applicazione di maschere facciali o nasali, oppure con modalità invasiva che comporta l'intubazione endotracheale. Quest'ultima può essere adottata per il trattamento di pazienti in condizioni critiche, del quale si occupa la medicina intensiva. Nella cura di questi pazienti intubati, lo svezzamento dal ventilatore meccanico è un elemento essenziale e universale.

Lo svezzamento copre l'intero processo di liberazione del paziente dal supporto meccanico e dal tubo endotracheale, compresi gli aspetti rilevanti della cura terminale (figura 1).



*Figura 1.* Lo svezzamento di un paziente dalla ventilazione meccanica invasiva si compie attraverso le seguenti fasi: 1)trattamento dell'insufficienza respiratoria acuta; 2)sospetto che lo svezzamento possa essere possibile; 3)valutazione della prontezza allo svezzamento; 4)prova di respirazione spontanea (SBT); 5)estubazione; e qualora questa fallisse 6)re-intubazione.

Trattata l'insufficienza respiratoria acuta, la seconda fase inizia la prima volta che il medico intuisce una ragionevole probabilità di successo nello svezzamento e si raggiunge, poi, la terza fase quando il medico esegue effettivamente dei test quotidiani di prontezza allo svezzamento, finché i risultati ottenuti giustificano una prova di respirazione spontanea (SBT). Questa ha generalmente una durata di 30 minuti, in quanto è stato dimostrato che la percentuale di successo di un SBT non varia prolungando la prova a

una durata di 120 minuti [2]. Valutando i parametri relativi al sistema respiratorio, l'adeguatezza dello scambio gassoso, la stabilità emodinamica e il comfort del paziente viene determinato il superamento del SBT. Una volta che il paziente dimostra l'abilità di sostenere una respirazione spontanea si valuta se il paziente potrebbe tollerare l'estubazione o necessitare di una re-intubazione. Fino al 25% dei pazienti che sperimentano un SBT con successo hanno poi bisogno di una re-intubazione nelle ore successive all'estubazione[3]. È una importante decisione siccome sia un'estubazione ritardata che una fallita sono associate a un aumento della durata della ventilazione e della mortalità.

Ogni nuovo giorno di ventilazione, senza che avvenga un'estubazione con successo, è associato a una mortalità progressivamente crescente. I pazienti non svezzati un giorno dopo il primo tentativo di estubazione presentano una mortalità del 19.0% mentre, per i pazienti ancora ricoverati in ICU e non estubati con successo dieci giorni dopo il primo tentativo, si arriva al 36.8% di mortalità [3]. Inoltre, una ventilazione prolungata aumenta la probabilità di contrarre infezioni polmonari associate al ventilatore e subire un trauma delle vie aeree, in quanto si ha un aumento dello stress sugli alveoli. Uno svezzamento ritardato comporta anche un aumento dei costi ospedalieri e delle risorse dell'unità di terapia intensiva utilizzate [4]. Un tentativo di estubazione prematuro, invece, espone il paziente a uno stress cardiovascolare e richiede spesso un ripristino del supporto ventilatorio [5]. È cruciale, perciò, determinare il momento appropriato per svezzare un paziente dalla ventilazione meccanica invasiva.

## 1.1 Scopo del progetto di tesi

Viste le problematiche sopracitate, emerge l'idea di sviluppare un sistema CAD (computer-aided diagnosis) per supportare il medico nell'individuazione del momento ottimale per svezzare il paziente dal ventilatore meccanico. Questo strumento potrebbe affiancare i già esistenti test quotidiani di prontezza allo svezzamento, dunque valutando giornalmente i parametri del paziente il sistema automatico potrebbe fornire una valutazione sull'eventuale fallimento o successo dell'estubazione. È importante evidenziare che uno strumento di aiuto alla decisione non sostituisce la diagnosi del medico esperto, ma fornisce delle opzioni che il decisore può accettare e condividere oppure no. Il progetto di tesi si pone, quindi, l'obiettivo di indagare i modelli di classificazione adatti al problema clinico in questione, esplorando gli esistenti studi. A questo scopo, è necessario valutare le variabili che descrivono al meglio le condizioni cliniche del paziente da svezzare, usufruendo del database dedicato alla medicina intensiva MIMIC-III (Medical Information Mart for Intensive Care).

## 1.2 Situazione attuale

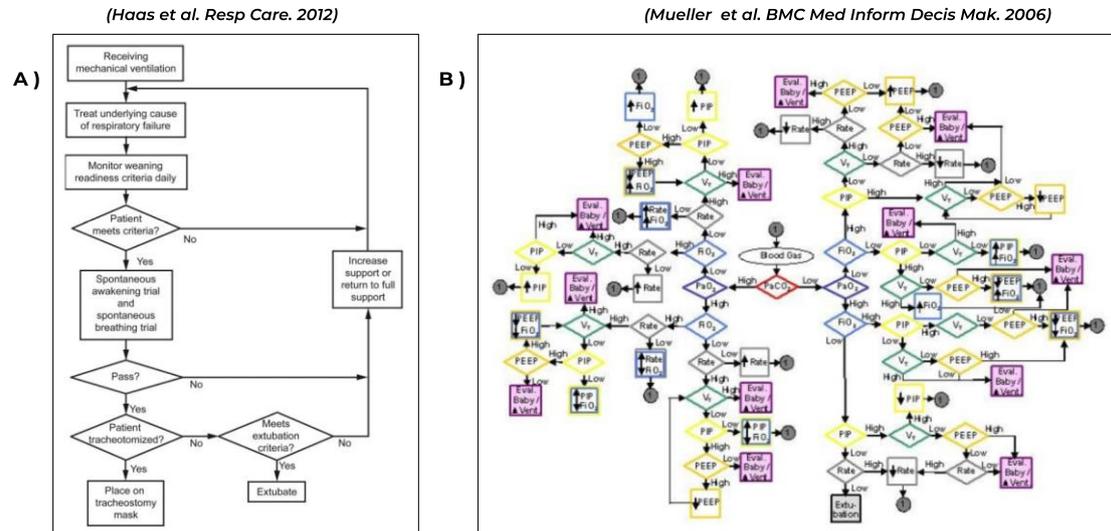
Ancora oggi vi è incertezza sui metodi migliori per condurre il processo di svezzamento e può accadere che diversi esperti arrivino a conclusioni differenti, in quanto hanno differenti livelli di esperienza e conoscenza. Al fine di ridurre queste differenze e dunque standardizzare l'approccio alla pratica clinica, a partire dagli anni 2000, sono state sviluppate linee guida, checklists e protocolli usando le evidenze ricavate da test e ricerche. Le linee guida sono istruzioni generali che permettono ai clinici di usare la propria esperienza e il proprio giudizio arrivando a decisioni anche differenti per lo stesso quadro clinico, mentre i protocolli contengono abbastanza dettagli da far sì che diversi esperti arrivino alla stessa decisione per uno stesso scenario clinico (figura 2 e figura 3). Le checklist aiutano, invece, ad assicurarsi che tutti gli aspetti della cura siano considerati tramite una lista di azioni e criteri da seguire [6]. La Evidence Based Medicine, dunque, garantisce che ogni paziente riceva il migliore trattamento possibile indipendentemente da chi gli fornisce le cure.

### ESEMPIO DI LINEE GUIDA (MacIntyre et al. Chest. 2001)

Evidence-based guidelines for weaning and discontinuing ventilatory support		
Number	Recommendation	Grade
1	In patients requiring mechanical ventilation > 24 hours, reversing all possible ventilatory and non-ventilatory issues should be an integral part of the ventilator discontinuance process.	B
2	A formal assessment of discontinuation potential should be made if the following criteria occur: <ul style="list-style-type: none"> <li>Evidence of reversal of the underlying cause for respiratory failure</li> <li>Adequate oxygenation (eg, <math>P_{aO_2}/F_{iO_2} &gt; 150</math>–200 mm Hg) on low PEEP (<math>\leq 5</math>–8 cm H<sub>2</sub>O) and pH <math>\geq 7.25</math></li> <li>Hemodynamic stability: no active myocardial ischemia, no clinically important hypotension (no vasopressors, low dose [<math>&lt; 5 \mu\text{g}/\text{kg}/\text{min}</math>] dopamine or dobutamine are OK)</li> <li>Capability to initiate an inspiratory effort</li> </ul>	B
3	Formal discontinuation assessment should be performed during a spontaneous breathing trial (SBT) rather than while receiving substantial ventilatory support. <ul style="list-style-type: none"> <li>An initial brief period of spontaneous breathing can be used to assess the capacity of continuing onto a formal SBT.</li> <li>The criteria for SBT tolerance include respiratory pattern, adequacy of gas exchange, hemodynamic stability, and subjective comfort.</li> <li>SBT tolerance lasting 30–120 min should prompt consideration for permanent ventilator discontinuation.</li> </ul>	A
4	After passing the SBT, artificial airway removal should be based on assessment of airway patency and ability of patient to protect the airway.	C
5	If failing an SBT, cause for failure should be determined. After reversing cause for failure, if patient meets SBT criteria, SBTs should be performed every 24 hours.	A
6	Patients failing an SBT should receive a stable, non-fatiguing, comfortable form of ventilatory support.	B
7	Anesthesia/sedation strategies and ventilator management aimed at early extubation should be used in postsurgical patients.	A
8	Discontinuation protocols designed for nonphysician healthcare providers should be developed and implemented by ICUs. Protocols aimed at optimizing sedation also should be developed and implemented.	A
9	Tracheostomy should be considered after an initial period of stabilization on the ventilator, when it becomes apparent the patient will require prolonged ventilator assistance. Patients who may benefit include those: <ul style="list-style-type: none"> <li>Requiring high levels of sedation to tolerate translaryngeal tubes</li> <li>Having marginal respiratory mechanics and reduced airway resistance from tracheostomy may reduce risk of muscle overload</li> <li>Potentially deriving psychological benefit from ability to eat orally, to communicate by speaking, and to experience enhanced mobility</li> <li>In whom enhanced mobility may assist physical therapy efforts</li> </ul>	B
10	Unless there is evidence for clearly irreversible disease (eg, high spinal cord injury, amyotrophic lateral sclerosis), a patient requiring prolonged ventilatory support should not be considered permanently ventilator-dependent until 3 months of weaning attempts have been made.	B
11	When medically stable, patients who have failed ventilator discontinuation attempts in the ICU should be transferred to a facility that specializes in such patients.	C
12	Weaning strategies in the prolonged mechanically ventilated patient should be slow-paced and should include gradually lengthening self-breathing trials.	C

**Figura 2.** Nel 2001 l'American College of Chest Physicians [7] sviluppa una linea guida relativa al processo di svezzamento dal ventilatore meccanico. Ad ogni raccomandazione è assegnata una votazione che indica il livello di affidabilità supportata dalle evidenze. Queste linee guida suggeriscono che una valutazione giornaliera del paziente è preferibile per determinare la prontezza di un paziente allo svezzamento e che i protocolli di svezzamento eseguiti da operatori sanitari, come infermieri, possono ridurre la durata della ventilazione e diminuire i costi associati [6].

## ESEMPI DI PROTOCOLLO



**Figura 3.** **A)** Diagramma che evidenzia i principali passaggi di un protocollo per lo svezzamento dalla ventilazione meccanica invasiva [6]. **B)** Diagramma di un algoritmo per la gestione della ventilazione meccanica adattato dallo studio di Carlo & Martin. *Pediatr Clin North Am.* 1986. Tale algoritmo mostra la complessità delle relazioni tra le variabili illustrando potenziali azioni (quadrati) che dovrebbero essere intraprese a seguito di determinati risultati delle emogasanalisi (al centro del diagramma) con l'obiettivo di raggiungere il riquadro che consiglia l'estubazione (al fondo del diagramma)[8].

Alcuni studi hanno effettuato un confronto tra processi di svezzamento condotti tramite l'applicazione di protocolli e processi di svezzamento guidati dal medico. Nel gruppo di pazienti che è stato trattato secondo un protocollo si è verificata una riduzione della durata di ventilazione, del numero di estubazioni fallite e dei costi ospedalieri. Inoltre, è stato notato che seguendo un protocollo, un infermiere anestesista riesce a eseguire un processo di svezzamento con performance equiparabili a quelle di un medico altamente specializzato. In altri studi, è emerso che adottando un protocollo di svezzamento si ottengono gli stessi risultati di un processo guidato da esperti che avviene in un reparto dove è presente un rapporto 1:1 tra pazienti e infermieri ed inoltre, il medico di terapia intensiva effettua una visita al letto del paziente due volte al giorno[6].

I protocolli di svezzamento prevedono, tra le altre, istruzioni concernenti la valutazione di alcuni parametri associati al paziente. Anche nella pratica clinica si ricorre all'analisi di indici, tra i quali il più comune è il Rapid Shallow Breathing Index (RSBI) che si calcola come rapporto tra la frequenza respiratoria e il volume corrente. In particolare, un paziente che presenta un RSBI superiore a 105 ha maggiori probabilità di andare incontro a fallimento dello svezzamento. Esistono vari indici che hanno lo scopo di stabilire l'esito del tentativo di estubazione, ma gli studi presenti in letteratura sull'uso di questi indici come parametri predittivi esprimono pareri contrastanti. Diversi studi concludono che gli indici valutati hanno un basso potere predittivo e non riescono a discriminare

efficientemente i pazienti che falliranno l'estubazione da quelli che la tollereranno. Alcuni studi suggeriscono che questi indici raggiungono un'accuratezza maggiore se combinati in unico indice integrato, invece che sfruttati singolarmente.

### **1.3 Passaggi seguiti nel progetto di tesi**

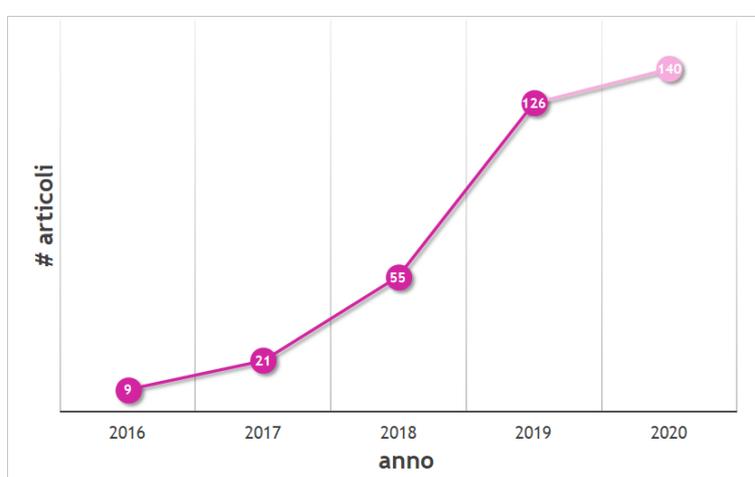
In seguito alla comprensione del problema clinico in questione e l'indagine su come altri studi lo hanno affrontato, il progetto di tesi ha seguito gli stessi passi dello sviluppo di un sistema CAD di supporto alla diagnosi. La prima fase della realizzazione del lavoro di tesi, dunque, è stata la costruzione del dataset e conseguente data cleaning. In una fase successiva, è stata effettuata una suddivisione del dataset creando, in particolare, un training set bilanciato e rappresentativo dell'intero dataset. Questi passaggi sono stati utili alla costruzione di due tipologie di classificatore, le Reti Neurali e il Light Gradient Boosting Machine (LightGBM), che verrà trattata nel sesto e settimo capitolo. Per la prima tipologia di modello è stata effettuata una fase di allenamento analizzando diverse architetture, una fase di testing e valutazione di queste e infine, una feature selection combinata all'ottimizzazione della struttura della rete neurale, tramite un algoritmo genetico. Affrontando la costruzione del LightGBM, invece, è stato effettuato l'allenamento del modello a seguito del tuning degli iperparametri, il testing e la validazione del classificatore interpretando, inoltre, i risultati del modello grazie all'importanza e il shap value e infine, una forward feature selection. Nel capitolo finale si confronteranno i risultati ottenuti con quelli presenti in letteratura e si discuteranno possibili sviluppi ed analisi future, al fine di proseguire lo studio sull'applicazione di metodi di data science e machine learning al problema clinico dello svezzamento dei pazienti intubati dal ventilatore meccanico.

## Capitolo 2

### Database MIMIC-III

Il database MIMIC-III (Medical Information Mart for Intensive Care) è un database relazionale creato dai ricercatori del laboratorio di fisiologia computazionale presso il MIT (Massachusetts Institute of Technology), con lo scopo di rendere i dati accessibili ai ricercatori internazionali e promuovere lo sviluppo e la riproducibilità di studi clinici (figura 4) [9]. Questo database è liberamente accessibile accettando un accordo sull'utilizzo dei dati.

**PUBBLICAZIONI CHE IMPIEGANO IL DATABASE MIMIC-III NEI LORO STUDI  
SECONDO IL DATABASE SCOPUS**



*Figura 4. La terza versione del database MIMIC-III è stata resa disponibile nell'anno 2016. Nella figura è mostrato il numero di pubblicazioni che impiegano il database MIMIC-III nei loro studi, nel corso degli anni fino al 2020.*

Il MIMIC-III raccoglie i dati anonimizzati di pazienti ricoverati nel reparto di terapia intensiva dell'ospedale Beth Israel Deaconess Medical Center a Boston, Massachusetts. Durante il periodo di raccolta, che si estende per circa una decade, sono stati collezionati dati associati a 53.423 ammissioni ospedaliere di pazienti adulti, in particolare negli anni tra il 2001 e il 2012 e dati associati a 7.870 ammissioni neonatali negli anni tra il 2001 e il 2008.

I dati presenti all'interno del MIMIC-III spaziano dai segnali vitali, i dettagli demografici, le medicazioni, gli esami di laboratorio alle note appuntate dai clinici, i referti del reparto di radiografia e i codici ICD-9 (International Classification of Diseases) per le procedure terapeutiche e le diagnosi.

Al fine di costruire il database, tutti i dati inclusi in esso hanno subito un processo di anonimizzazione, in conformità allo standard HIPAA (Health Insurance Portability and Accountability Act) per garantire la protezione dei dati sensibili dei pazienti.

I regolamenti HIPAA sono stati sviluppati negli Stati Uniti con scopo sostanzialmente analogo a quello del GDPR (General Data Protection Regulation) in vigore nell'Unione Europea. Sono stati, quindi, rimossi dai creatori del database tutti quei campi come il nome del paziente, il recapito telefonico, l'indirizzo email, il codice postale e il numero di previdenza sociale. Inoltre, per evitare di risalire all'identità dei pazienti le indicazioni temporali sono state traslate nel futuro di una quantità casuale per ogni paziente, mantenendo però inalterati gli intervalli temporali e rendendo tutte le ammissioni ospedaliere come avvenute tra l'anno 2100 e l'anno 2200.

Il database MIMIC-III è di tipo relazionale e i suoi dati sono, dunque, organizzati in 26 tabelle collegate tra loro. Ogni riga di una tabella corrisponde a un record identificato univocamente da un ID, ovvero una chiave. Le colonne della tabella contengono, invece, gli attributi dei records. Avvalendosi delle associazioni tra le chiavi è possibile usufruire delle relazioni tra le tabelle. Alcune delle chiavi più rilevanti nello studio di pazienti intubati in condizioni critiche sono il SUBJECT\_ID che identifica univocamente i pazienti, il HADM\_ID che si riferisce a un'unica ammissione ospedaliera e il ICUSTAY\_ID che identifica, invece, un'unica ammissione in terapia intensiva. È importante mettere in evidenza che un paziente può essere associato a più di un'ammissione ospedaliera, in quanto, negli anni dedicati alla raccolta dei dati da inserire nel database il paziente può essere stato ricoverato più volte. Inoltre, durante una singola ammissione ospedaliera, un paziente può anche subire uno o molteplici ricoveri nel reparto di terapia intensiva. Nella figura 5 è visibile un'anteprima della tabella ICUSTAYS, che insieme alle tabelle ADMISSIONS, PATIENTS, SERVICES e TRANSFERS definisce e traccia le degenze dei pazienti.

ANTEPRIMA TABELLA ICUSTAYS

Riga	ROW_ID	SUBJECT_ID	HADM_ID	ICUSTAY_ID	DBSOURCE	FIRST_CAREUNIT	LAST_CAREUNIT	FIRST_WARDID	LAST_WARDID	INTIME	OUTTIME	LOS
1	2	3	145834	211552	carevue	MICU	MICU	12	12	2101-10-20T19:10:11	2101-10-26T20:43:09	6.0646
2	8	8	159514	262299	carevue	NICU	NICU	56	56	2117-11-20T12:36:10	2117-11-21T14:24:55	1.0755
3	9	9	150750	220597	carevue	MICU	MICU	15	15	2149-11-09T13:07:02	2149-11-14T20:52:14	5.3231
4	12	12	112213	232669	carevue	SICU	SICU	23	23	2104-08-08T02:08:17	2104-08-15T17:22:25	7.6348
5	13	13	143045	263738	carevue	CCU	CSRU	57	14	2167-01-08T18:44:25	2167-01-12T10:43:31	3.666
6	15	17	194023	277042	carevue	CSRU	CSRU	12	12	2134-12-27T16:21:48	2134-12-29T18:04:03	2.071
7	16	17	161087	257980	carevue	CSRU	CSRU	12	12	2135-05-09T14:12:06	2135-05-10T11:18:34	0.8795
8	19	20	157681	264490	carevue	CSRU	CSRU	14	14	2183-04-28T15:00:36	2183-04-29T16:13:48	1.0508
9	22	22	165315	204798	carevue	MICU	MICU	52	52	2196-04-09T12:27:00	2196-04-10T15:54:00	1.1438
10	23	23	152223	227807	carevue	CSRU	CSRU	14	14	2153-09-03T09:38:55	2153-09-04T15:59:11	1.2641
11	29	28	162569	225559	carevue	CSRU	CSRU	14	14	2177-09-01T09:32:26	2177-09-02T12:28:42	1.1224
12	31	31	128652	254478	carevue	MICU	MICU	15	15	2108-08-22T23:28:42	2108-08-30T21:59:20	7.9379
13	32	32	175413	295037	carevue	TSICU	TSICU	33	33	2170-04-04T18:46:00	2170-04-12T11:13:00	7.6854

Figura 5. Anteprima della tabella ICUSTAYS le cui righe sono disposte in ordine crescente secondo il codice univoco dei pazienti. Si può notare (in giallo) la molteplicità di ricoveri in terapia intensiva che uno stesso paziente può subire.

Alcune tabelle, come CHARTEVENTS, LABEVENTS, INPUTEVENTS\_CV e INPUTEVENTS\_MV contengono dati associati alla cura e il trattamento dei pazienti e raccolgono quindi parametri relativi a impostazioni ventilatorie, analisi ematiche, emogasanalisi, analisi delle urine, segnali vitali e altre misurazioni fisiologiche. Nel database sono presenti dati registrati tramite due differenti sistemi informativi: CareVue e MetaVision. Ove possibile

i dati sono stati aggregati, ed il riferimento al sistema da cui provengono è stato mantenuto aggiungendo un attributo apposito alla descrizione delle variabili registrate. Per quelle registrazioni che presentano, però, differenze significanti nella struttura degli attributi, cioè dei campi della tabella, sono state conservate le tabelle originali distinte. Ne sono un esempio quelle relative alle somministrazioni di farmaci: INPUTEVENTS\_CV e INPUTEVENTS\_MV.

Altre cinque tabelle D\_CPT, D\_ICD\_DIAGNOSES, D\_ICD\_PROCEDURES, D\_ITEMS e D\_LABITEMS fungono da dizionari per consultare le definizioni e descrizioni dei codici identificativi dei parametri presenti nelle varie tabelle.

## Capitolo 3

### Costruzione del dataset

In seguito all'analisi del problema clinico che si vuole affrontare, la prima fase dello sviluppo di un sistema CAD consiste nella costruzione del dataset, dunque nell'individuazione di una fonte da cui estrarre i dati, nella selezione dei soggetti da analizzare secondo criteri di inclusione ed esclusione e nella definizione dei parametri che costituiranno le features caratterizzanti ogni paziente.

Il database MIMIC-III è stato individuato come fonte di parametri associati a pazienti intubati in condizioni critiche, in quanto è una vasta raccolta di informazioni che si estende per un periodo di dodici anni e comprende un elevato numero di pazienti. Al fine di costruire il dataset da studiare si è intrapresa un'elaborazione delle tabelle relazionali, tramite l'implementazione di query sulla piattaforma di Google Cloud BigQuery.

Il dataset, infine ottenuto, è costituito da un totale di 13691 pazienti, e dunque da un egual numero di eventi di ventilazione meccanica invasiva, e 85 variabili che li descrivono.

#### 3.1 Selezione dei soggetti

Per studiare il problema clinico dello svezzamento dal ventilatore meccanico è necessario definire la popolazione target, rappresentata in questa analisi dai pazienti adulti sottoposti a un trattamento di ventilazione meccanica invasiva. Il campione di soggetti disponibili per la ricerca è ricavato dal database MIMIC-III estraendo tutti quei pazienti ricoverati nell'unità di terapia intensiva, dell'ospedale Beth Israel Deaconess Medical Center a Boston, che hanno subito l'intubazione endotracheale negli anni tra il 2001 e il 2012. Ai criteri di inclusione, quali la maggiore età, la condizione clinica, l'indicazione geografica e temporale si aggiungono dei criteri di esclusione tramite i quali si arriva ad avere la popolazione di soggetti effettivamente reclutati.

##### 3.1.1 Estrazione degli eventi di ventilazione meccanica invasiva

Il primo passo necessario alla selezione dei soggetti è stato la determinazione dello stato di 'intubato' / 'non intubato', analizzando i parametri di ventilazione per ogni ricovero di ogni paziente contenuto nel database, in quanto in esso non è presente un'informazione diretta che esplicita quando il paziente è intubato oppure no. Se il paziente presenta dei parametri relativi all'uso di un ventilatore meccanico è considerato intubato, quindi la prima presenza di un parametro di ventilazione indica l'inizio di un nuovo evento di ventilazione meccanica invasiva. Inoltre, la presenza di un parametro di ventilazione nelle 8 ore seguenti continua l'evento, aggregando le impostazioni di ventilazione consecutive come un singolo evento di ventilazione. L'evento di ventilazione corrente viene interrotto se si riscontra un'estubazione documentata o l'inizio di una ventilazione non invasiva o di una terapia con ossigeno. Infine, il punto in cui si passa dallo stato 'intubato' a quello 'non intubato' viene considerato come l'evento di estubazione. Gli eventi di ventilazione così estratti devono essere poi ricondotti al codice identificativo del paziente, siccome

sono direttamente associati al solo codice identificativo del ricovero in terapia intensiva. Combinando, allora, le informazioni presenti nella tabella ICUSTAYS è possibile risalire al paziente coinvolto nel ricovero in cui ha avuto luogo l'intubazione.

In questo modo, è stato possibile identificare i pazienti sottoposti a intubazione, tra la totalità dei pazienti ricoverati in terapia intensiva. Anche in questo caso, è importante evidenziare che durante uno stesso ricovero in ICU un paziente può essere stato sottoposto a molteplici trattamenti invasivi di supporto ventilatorio, dovuti anche eventualmente al fallimento di precedenti tentativi di estubazione. È stata implementata, perciò, una query che ha permesso di ottenere una nuova tabella contenente tutti gli eventi di ventilazione meccanica invasiva presenti nel database associando a ognuno il tempo di inizio e fine, la durata e l'ordine di accadimento.

ANTEPRIMA TABELLA VENTDURATION

Riga	icustay_id	ventnum	starttime	endtime	duration_hours
101	200045	8	2116-07-29T12:40:00	2116-07-30T07:45:00	19
102	200045	9	2116-07-30T09:00:00	2116-07-30T10:02:00	1
103	200045	10	2116-07-30T11:00:00	2116-07-30T14:00:00	3
104	200046	1	2154-05-01T15:00:00	2154-05-11T00:00:00	225
105	200046	2	2154-05-11T01:00:00	2154-06-06T09:00:00	632
106	200046	3	2154-06-06T10:00:00	2154-07-26T17:00:00	1207
107	200053	1	2166-02-27T18:00:00	2166-03-04T04:00:00	106
108	200059	1	2198-02-10T06:00:00	2198-02-17T14:30:00	176
109	200059	2	2198-02-17T15:00:00	2198-02-17T19:00:00	4
110	200059	3	2198-02-17T20:30:00	2198-02-18T21:00:00	25
111	200059	4	2198-02-18T22:35:00	2198-02-19T23:00:00	25
112	200059	5	2198-02-20T01:30:00	2198-02-20T12:00:00	11
113	200059	6	2198-02-20T13:27:00	2198-02-23T11:00:00	70

Figura 6. Anteprima della tabella VENTDURATION le cui righe sono disposte in ordine crescente secondo il codice identificativo dell'ammissione nel reparto di terapia intensiva. Si può notare (in giallo) la molteplicità di eventi di ventilazione meccanica invasiva a cui un paziente può essere sottoposto durante un unico ricovero in ICU.

Tramite queste operazioni, sono stati ottenuti 51575 eventi di ventilazione meccanica invasiva, avvenuti durante 27912 distinte ammissioni in terapia intensiva corrispondenti a 23968 pazienti che sono stati intubati. Questa assenza di corrispondenza biunivoca tra l'insieme dei pazienti, dei ricoveri e delle intubazioni è dovuta alla molteplicità di ricoveri e intubazioni che un singolo paziente può subire.

### 3.1.2 Scelta e applicazione dei criteri di inclusione ed esclusione

Il campione di soggetti definito dai criteri di inclusione, è stato ulteriormente delimitato con l'applicazione di criteri di esclusione, in modo tale da massimizzare la validità dei risultati nel contesto specifico di questo studio. Infatti, tali criteri sono stati scelti per descrivere più accuratamente i pazienti da studiare, le quali caratteristiche saranno le stesse dei pazienti ai quali si applicherà lo strumento sviluppato e quindi questi criteri permettono di migliorare la generalizzabilità dei risultati dello studio.

I criteri di esclusione individuati per evitare di inserire nel dataset pazienti che potrebbero interferire con la qualità dei risultati dello studio sono spiegati di seguito.

- **Criterio dell'età**

Il database contiene dati appartenenti a pazienti adulti (con età maggiore o uguale ai 15 anni) e neonati (con età compresa tra 0 e 1 anno); non sono invece presenti dati appartenenti a pazienti pediatrici (con età maggiore a 1 anno e inferiore a 15 anni). È possibile calcolare l'età dei pazienti, al momento dell'ammissione in ospedale, facendo la differenza tra la data di ricovero (attributo *intime*) presente nella tabella ICUSTAYS e la data di nascita dei pazienti (attributo DOB) presente nella tabella PATIENTS.

In base ai criteri di inclusione sono stati selezionati solamente quegli eventi di ventilazione corrispondenti a pazienti adulti, cioè con età maggiore o uguale ai 18 anni. In conformità con la regolamentazione HIPAA, i creatori del database hanno traslato, di una quantità casuale, le date di nascita di tutti i pazienti con più di 89 anni, in modo tale che non si possa risalire alla loro vera età nel momento del ricovero. Questi pazienti appaiono, quindi, nel database con una età di oltre 300 anni. Siccome non è possibile per tali soggetti, estrapolare l'informazione relativa all'età è stato scelto di escludere gli eventi di ventilazione associati a pazienti con età maggiore di 89 anni. Rispetto al totale di 51575 eventi di ventilazione meccanica invasiva associati a 23968 distinti pazienti, in seguito all'applicazione dei criteri sull'età restano 47224 eventi di ventilazione corrispondenti a 21244 differenti pazienti.

- **Criterio della morte avvenuta prima del tentativo di estubazione**

Si vogliono escludere le ventilazioni nelle quali il paziente è morto prima del tentativo di estubazione, in quanto non necessitando di una successiva re-intubazione verrebbero classificate erroneamente come estubazioni riuscite o altrimenti non è possibile classificarle come estubazioni fallite, siccome non è mai avvenuto il tentativo di estubazione. Nella tabella PATIENTS del database è presente l'informazione relativa alla data di morte (attributo DOD) e allora è possibile calcolare la differenza in secondi tra questa e la data di fine ventilazione, (attributo *endtime*) nella tabella VENTDURATION. Siccome nel database le data di morte dei pazienti risultano sempre avvenuta all'orario 00.00 sono state escluse tutte quelle ventilazioni per le quali la differenza calcolata è negativa o nulla, quindi tutte quelle ventilazioni nelle quali il paziente è deceduto lo stesso giorno dell'estubazione o prima del tentativo di estubazione.

- **Criterio della durata**

Al fine di escludere le intubazioni di routine a seguito di interventi chirurgici, il cui processo di svezzamento è a basso rischio di fallimento sono stati esclusi tutti quegli eventi di ventilazione con durata inferiore a 24 ore [10,11]. Inoltre, tali eventi non rientrano nella definizione dei pazienti in condizioni critiche, protagonisti del problema clinico dello svezzamento descritto.

- **Criterio dell'estubazione non pianificata**

Durante la ventilazione meccanica invasiva può avvenire un'estubazione non programmata del paziente, cioè non desiderata dallo staff medico. Può accadere

che questa sia accidentale, e quindi provocata da operatori sanitari durante l'esecuzione di interventi sul paziente, oppure può accadere che il paziente rimuova volontariamente il tubo endotracheale, cioè avviene un'auto-estubazione. L'incidenza di questo fenomeno nei pazienti adulti, sottoposti al supporto ventilatorio invasivo, varia dal 2 al 16% [12]. È stato deciso di non inserire nel dataset le ventilazioni nelle quali è avvenuta una estubazione non programmata, in modo tale da non creare ulteriore variabilità tra le caratteristiche dei casi di ventilazione analizzati.

- **Criterio del primo evento di ventilazione meccanica invasiva**

Ogni paziente può essere sottoposto a molteplici trattamenti di ventilazione meccanica invasiva durante uno stesso ricovero, dovuti al fallimento dei tentativi di svezzamento. È stato deciso di escludere tutte quelle ventilazioni successive alla prima, siccome i pazienti a seguito di un'estubazione fallita o in condizione di ventilazione prolungata presentano uno scenario clinico differente e comporterebbero una disomogeneità dei casi di ventilazione esaminati.

- **Criterio del primo ricovero in terapia intensiva per ogni paziente**

Ogni paziente può aver subito molteplici ricoveri nel reparto di terapia intensiva durante il periodo di raccolta dei dati per la costruzione del database. Se si mantenesse più di un ricovero per uno stesso paziente, gli eventi di ventilazione a essi associati saranno altamente correlati appartenendo allo stesso paziente. Si sceglie, dunque, di escludere tutte le ammissioni in ICU successive alla prima per un paziente. In questo modo si semplificano le analisi successivamente effettuate, avendo elementi indipendenti tra loro e si ottiene un solo evento di ventilazione per ogni paziente.

A seguito dell'applicazione dei criteri descritti si ottiene un insieme di 13691 pazienti a cui è associata una singola ammissione nel reparto di terapia intensiva ed un singolo evento di ventilazione meccanica invasiva.

### 3.2 Determinazione della classe di ogni evento di ventilazione

Non sempre un paziente che sperimenta una prova di respirazione spontanea con esito favorevole riesce poi a sostenere la funzione respiratoria autonomamente, portando quindi al fallimento dello svezzamento. Questo è definito nella maggior parte degli studi dalla necessità di reinstaurare il supporto ventilatorio, intubando nuovamente il paziente, entro 48 ore dal tentativo di estubazione.

Ogni evento di ventilazione estratto dal database è stato classificato in due categorie: estubazione riuscita (ES), considerata in seguito per la costruzione del classificatore come classe positiva ed estubazione fallita (EF), considerata come classe negativa. Sfruttando i campi della tabella VENTDURATION, in particolare la numerazione degli eventi di ventilazione durante uno stesso ricovero in terapia intensiva e le date di inizio e fine della ventilazione meccanica invasiva, è stato possibile calcolare l'intervallo in ore che intercorre tra un'estubazione e l'eventuale re-intubazione successiva a essa. Tutti quegli eventi di ventilazione seguiti da una re-intubazione nelle 48 ore successive all'estubazione sono stati inseriti nella classe delle estubazioni fallite, mentre gli altri sono

stati definiti come estubazioni riuscite. Applicando questo metodo, i 13691 eventi di ventilazione, che costituiranno i campioni del dataset, risultano suddivisi in 10835 estubazioni riuscite (79,14% delle ventilazioni totali) e 2856 estubazioni fallite (20,86% delle ventilazioni totali). La percentuale di estubazioni fallite è in linea con quella riportata negli studi che trattano lo svezzamento dalla ventilazione meccanica invasiva, i quali riportano un dato che può arrivare anche al fallimento dello svezzamento in un paziente su quattro.

### 3.3 Calcolo ed estrazione delle variabili

Al fine di determinare se un paziente intubato in condizioni critiche, in un determinato momento, tollererebbe lo svezzamento è necessario valutare diversi parametri che descrivono lo scenario clinico del paziente, la facoltà respiratoria e i parametri ventilatori e, dunque, caratterizzare il paziente e l'esperienza di ventilazione meccanica invasiva che sta affrontando. È stata, perciò, eseguita un'indagine sulle variabili scelte e utilizzate in altri studi che trattano il problema clinico dello svezzamento dalla ventilazione meccanica o che analizzano la stessa popolazione target coinvolta nello studio attuato dal presente progetto di tesi (Appendice A). In seguito a questa ricerca sono state definite 91 variabili da calcolare o estrarre dal database MIMIC-III, valutando anche l'effettiva disponibilità dei parametri raccolti nel database e la concreta possibilità di calcolo e derivazione di alcuni parametri a partire da quelli esistenti.

Le variabili scelte sono distinguibili in 7 macrocategorie: dati demografici, punteggi che valutano la gravità di pazienti in condizioni critiche, comorbidità, somministrazione sedativi e vasopressori, parametri caratterizzanti la funzionalità respiratoria, segnali vitali, analisi ematiche ed emogasanalisi.

I dati demografici dei pazienti sono, in particolare, l'età, il sesso, il peso registrato il giorno dell'intubazione, la variazione di peso tra il giorno dell'intubazione e quello dell'estubazione e l'altezza.

Le scale di punteggio, che valutano la gravità dello stato di pazienti in condizioni critiche, inserite nel dataset sono il SOFA score (Sequential Organ Failure Assessment) e il GCS score (Glasgow Coma Scale). Il SOFA score ha l'obiettivo di fornire una misura oggettiva della gravità della disfunzione degli organi di un paziente. Il punteggio totale è cumulativo di sei punteggi che valutano il sistema respiratorio, la funzionalità renale, il sistema cardiovascolare, l'organo epatico, il sistema nervoso centrale e la coagulazione. Il SOFA score può variare in un range di valori discreti tra 0 (funzione normale) e 24 (alto livello di disfunzione). Il GCS score fornisce un metodo pratico per la valutazione del livello di coscienza di un paziente in risposta a stimoli definiti. Anch'esso è composto da tre punteggi che valutano tre diverse funzioni neurologiche: l'apertura degli occhi, la risposta motoria e la risposta verbale a determinati stimoli o comandi. La scala di Glasgow può variare in un range di valori discreti che va da 3 (completa incoscienza) a 15 (coscienza normale). Entrambi i punteggi, aggiunti come variabili nel dataset, sono stati calcolati al primo giorno di ammissione in terapia intensiva.

Nel dataset è stata inserita una variabile che indica il numero di giorni in cui il paziente è stato sottoposto a sedazione. Sono stati, quindi, calcolati i giorni che intercorrono tra la

prima somministrazione di un sedativo (a partire dal primo giorno di ricovero in ICU) e l'ultima somministrazione (al massimo corrispondente al momento dell'estubazione). Una seconda variabile binaria indica l'avvenuta somministrazione durante la ventilazione meccanica invasiva di farmaci vasopressori.

Allo scopo di inserire informazioni relative alle patologie coesistenti dei pazienti sono state create una serie di variabili binarie che indicano la presenza o assenza di 29 patologie. Queste informazioni sono state tratte dalla tabella DIAGNOSES\_ICD presente nel database MIMIC-III. Nelle tabelle DIAGNOSES\_ICD e PROCEDURES\_ICD del database sono raccolti i codici ICD-9 associati a ogni ammissione ospedaliera, che sono stati registrati dall'ospedale al momento della dimissione a scopi amministrativi, come generalmente avviene. Il sistema di codifica ICD (International Classification of Disease) ha l'obiettivo di standardizzare la classificazione delle malattie, i disturbi, le lesioni e altre condizioni cliniche. La nona versione del codice è stata approvata dall'Organizzazione Mondiale della Sanità nel 1975 ed in seguito, è stata sviluppata una versione ampliata che introduce la classificazione degli interventi chirurgici e delle procedure diagnostiche e terapeutiche così da poter tenere traccia, nella scheda di dimissione ospedaliera, di quelle erogate. Negli Stati Uniti è in uso dal 2015 la decima versione della codifica, ma è già stata presentata l'undicesima versione che entrerà in vigore negli stati membri dell'organizzazione mondiale della sanità dal 2022.

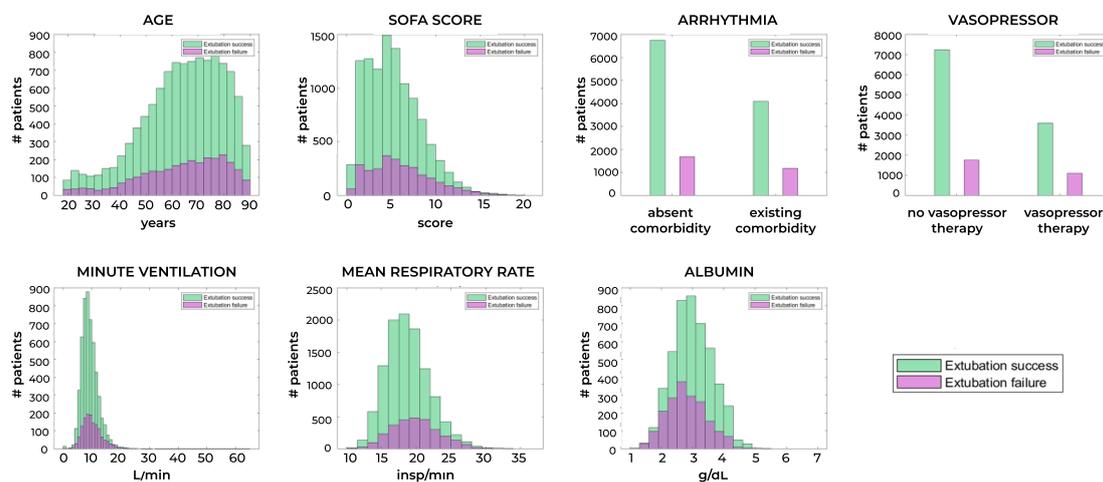
Nella costruzione del dataset sono stati inseriti anche cinque parametri che descrivono la funzione respiratoria e il supporto ventilatorio. In particolare, è stata inserita una variabile che indica la durata in ore della ventilazione meccanica e altre quattro dedicate all'ultimo valore registrato prima dell'estubazione dei parametri di volume corrente, volume ventilatorio al minuto, la pressione media delle vie aeree e la massima pressione inspiratoria.

I segnali vitali considerati sono la frequenza cardiaca e respiratoria, la temperatura corporea, la pressione sanguigna sistolica e diastolica, la pressione media arteriosa e la saturazione dell'ossigeno nel sangue. Durante il periodo di raccolta dei dati per la costruzione del database MIMIC-III sono state registrate le misurazioni dei parametri vitali, effettuate ogni ora dal personale infermieristico. Per ognuno dei segnali considerati sono state create tre variabili che indicano il valore medio, minimo e massimo del segnale calcolati durante l'intera ventilazione meccanica.

Infine, sono state estratte ulteriori variabili che indicano l'ultimo valore registrato prima dell'estubazione di tutti quei valori esaminati durante le analisi di laboratorio dei comuni prelievi sanguigni e dei prelievi di sangue arterioso. I primi, effettuati su un campione di sangue venoso, valutano valori come la concentrazione di emoglobina, globuli bianchi, piastrine, elettroliti, proteine e glucosio nel sangue. Grazie a questi esami è stato possibile inserire anche valori che valutano la coagulazione, la funzionalità renale, la funzionalità epatica e il rischio di insufficienza cardiaca. L'emogasanalisi ha, invece, proprio lo scopo di valutare la funzionalità polmonare e l'eventuale presenza di uno squilibrio metabolico, respiratorio o renale. Sono state, quindi, inserite variabili relative al pH arterioso, alla pressione parziale di ossigeno e anidride carbonica nel sangue, all'eccesso di basi nel sangue.

Tutte le 91 variabili calcolate e derivate dalle tabelle del database relazionale (Appendice B) sono state analizzate in termini di distribuzione separando i campioni nelle due classi. Nella figura 7 sono mostrate alcune rappresentazioni esemplificative.

#### RAPPRESENTAZIONE DELLA DISTRIBUZIONE DI ALCUNE VARIABILI NELLE DUE CLASSI (ES,EF)



*Figura 7. Rappresentazione della distribuzione di alcune variabili, una per ciascuna delle 7 macrocategorie individuate, nelle due classi in cui sono stati distinti i pazienti del dataset. È possibile notare la presenza di variabili di tipo binario (come ad esempio Arrhythmia e Vasopressor), di tipo discreto (come ad esempio SOFA score) e di tipo continuo (come ad esempio Age, Minute ventilation, mean Respiratory Rate e Albumin).*

## Capitolo 4

### Data cleaning

La fase successiva all'estrazione delle variabili è rappresentata dal processo di data cleaning. Esso prevede la gestione degli outliers, delle inconsistenze, dei valori mancanti e l'eventuale eliminazione di alcune variabili al fine di migliorare la qualità dei dati e della successiva estrazione di conoscenza da essi.

In primo luogo è stata valutata la percentuale di valori mancanti per ognuna delle 91 variabili. Da questa analisi è risultato che 35 variabili sono complete dell'informazione per tutti i pazienti presenti nel dataset, 34 variabili mostrano una percentuale di valori mancanti inferiore al 1%, 13 variabili hanno una percentuale di valori mancanti compresa tra il 2 e il 40% ed infine 9 variabili che presentano una percentuale superiore al 40%.

L'analisi sui valori mancanti ha portato alla luce che per quanto riguarda le variabili relative alla frequenza cardiaca, respiratoria e saturazione di ossigeno, alle quali manca il valore associato a un solo paziente, questo è sempre lo stesso. Inoltre, per le variabili relative alla pressione sistolica, diastolica e pressione arteriosa media, alle quali manca il valore associato a tre pazienti, questi sono sempre gli stessi tre, tra cui anche il paziente a cui si è fatto precedentemente riferimento. Dunque, è stato deciso di eliminare questi tre pazienti che presentavano un elevato numero di variabili mancanti, azzerando quindi il grado di valori mancanti per le diciotto variabili coinvolte.

Un'ulteriore decisione è stata quella di non considerare nella composizione del dataset finale le variabili con una quantità di valori mancanti superiore al 70% dei pazienti considerati, che sono nello specifico la concentrazione nel sangue di proteine totali (missing degree 96,7%), la concentrazione del peptide natriuretico cerebrale (missing degree 96,7%), la concentrazione della proteina C reattiva (missing degree 98%) e la concentrazione di troponina (missing degree 74,8%). Per le variabili altezza (missing degree 32,3%) e peso (missing degree 17,4%) è stato preferito escludere anch'esse dalla composizione finale del dataset piuttosto che intraprendere una sostituzione dei loro valori mancanti, in quanto è stata ritenuta discutibile la relazione dei valori di altezza e peso di un paziente con quelli delle altre variabili presenti.

In seguito a queste operazioni il numero di features realmente usato per la costituzione del dataset si è ridotto a 85.

Il processo di data cleaning è stato proseguito intraprendendo l'imputazione degli elementi mancanti relativi a 32 variabili. Per ognuna di queste si sono confrontati diversi metodi di imputazione, in quanto la sostituzione dei valori mancanti in una variabile può portare all'introduzione di un bias più o meno evidente nella distribuzione dei valori di quella variabile.

Il primo metodo applicato consiste nella sostituzione dei valori mancanti di una variabile con il valore medio (nel caso di una variabile continua) o mediano (nel caso di una variabile discreta) di tutti i valori presenti associati ai restanti pazienti per quella stessa variabile. È importante sottolineare che in questa operazione si è lavorato sulle due classi, estubazione riuscita e fallita, separatamente. Dunque, i valori mancanti dei pazienti di

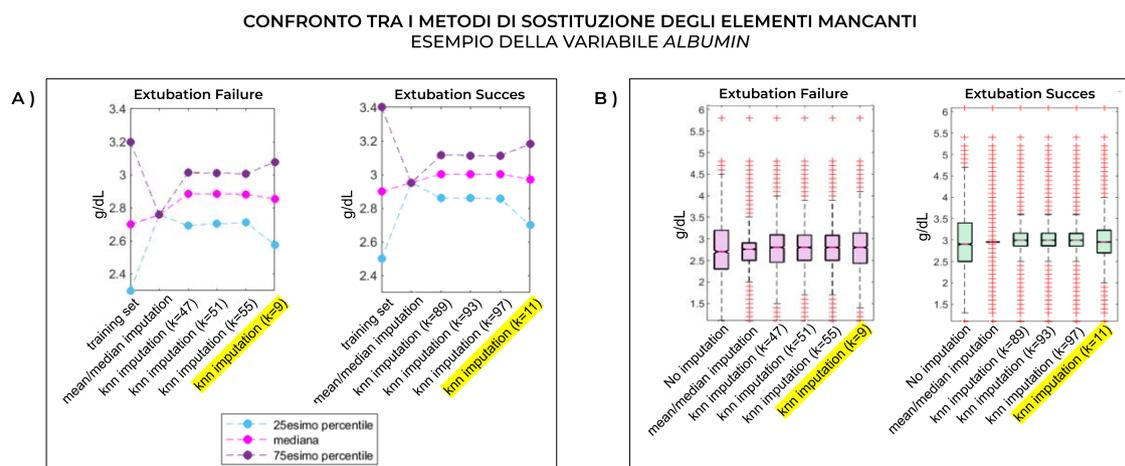
classe EF per una variabile continua sono stati sostituiti con il valore medio di quella variabile considerando solamente gli altri pazienti della stessa classe e allo stesso modo è stata eseguita l'imputazione per la classe ES.

Il secondo metodo applicato è stata l'imputazione dei valori mancanti tramite l'algoritmo kNN (k-nearest neighbor) che si basa sul calcolo della vicinanza di un elemento rispetto a un insieme di altri elementi. La metrica utilizzata in questo progetto di tesi per calcolare la vicinanza è stata la distanza euclidea. Anche per questo metodo sono state considerate separatamente le due classi, dunque l'imputazione di valori appartenenti a pazienti in una certa classe vede coinvolti solamente altri paziente di quella stessa classe. Per ogni variabile è stato considerato un insieme di pazienti tra cui cercare i più vicini a quello che presenta il valore mancante. Un paziente è descritto da una serie di attributi, cioè un vettore, e per calcolare la similarità tra due pazienti viene calcolata la distanza euclidea tra i due vettori a essi associati. Per effettuare il processo di imputazione sono state considerate solamente alcune tra tutte le variabili del dataset, in particolare solo quelle che presentavano un missing degree inferiore al 20%. Si evidenzia che riferendosi a tutte le variabili del dataset si intendono tutte quelle rimaste a seguito della rimozione delle sei variabili precedentemente descritte. È stato ottenuto quindi un totale di 78 attributi descrittivi ogni paziente e un totale di 11322 pazienti che posseggono il valore di tutte le 78 variabili. Questi pazienti saranno poi ripartiti in due sottogruppi separando quelli della classe delle estubazioni fallite da quelli della classe delle estubazioni riuscite, che conterranno rispettivamente 2587 pazienti e 8735 pazienti. In questo modo il sottoinsieme di pazienti, che verrà utilizzato per ricercare quelli più simili a ogni paziente che avrà un valore da imputare, non risulta eccessivamente ridotto dal grado di missing values di alcune variabili. A questo punto per ogni paziente che ha un valore mancante in una certa variabile dovrà cercare i k pazienti a lui più simili, considerando per il calcolo della distanza tra i pazienti tutti gli attributi tranne quello della variabile che si vuole imputare, essendo ovviamente mancante nel paziente in questione. In questo algoritmo è essenziale la determinazione del numero k di pazienti più vicini che verranno cercati tra il sottoinsieme di ricerca a disposizione. Generalmente è consigliato utilizzare un valore di k pari alla radice quadrata del numero di elementi presenti nel training set, corrispondente in questo caso a 93 per la classe delle estubazioni riuscite e 51 per la classe delle estubazioni fallite.

Trovati i k pazienti più vicini a quello che si vuole imputare viene calcolata la media (o mediana se è una variabile discreta) dei valori, appartenenti ai pazienti del training set, della variabile in cui il paziente in questione presenta il missing value. Questo viene infine sostituito con il valor medio (o mediano) così calcolato. L'intera procedura è stata effettuata valutando diversi valori di k. Per la classe ES, che è la più numerosa, sono stati provati valori di k pari a 89, 93, 97 e 11 mentre per la classe EF pari a 47, 51, 55 e 9.

Per quanto riguarda le 7 variabili non considerate in questa procedura, ne è stata eseguita una analoga, tranne che per la costituzione del training set. Infatti, per ognuna di queste variabili è stato utilizzato un training set ricavato considerando le 78 variabili precedenti ed in più la variabile che si vuole imputare. Ottenendo ogni volta un totale di 79 variabili sulla base delle quali sarà estratto il sottoinsieme dei pazienti che posseggono il valore

per tutte le 79 variabili. In questo modo il numero di elementi nel training set, anche se inevitabilmente inferiore al caso precedente, sarà comunque non eccessivamente scarso. La valutazione dei diversi valori del parametro k provati è stata eseguita valutando l'entità del bias introdotto nella distribuzione delle variabili imputate tramite il processo di sostituzione dei valori mancanti. È stata quindi valutata la differenza nella distribuzione dei valori di una variabile prima e dopo l'esecuzione di tale processo, ponendo attenzione al valore mediano, al 75° percentile e al 25° percentile della distribuzione (figura 8).



**Figura 8. A)** Nella rappresentazione è possibile confrontare la distribuzione per la variabile 'Albumin' dei pazienti che hanno il valore per tale variabile (training set) con la distribuzione dei soli valori sostituiti tramite i metodi di sostituzione indicati per quei pazienti che non possedevano il valore per tale variabile. **B)** La raffigurazione dei boxplot permette di confrontare la distribuzione della variabile prima del processo di sostituzione degli elementi mancanti e a seguito del processo tramite i diversi metodi.

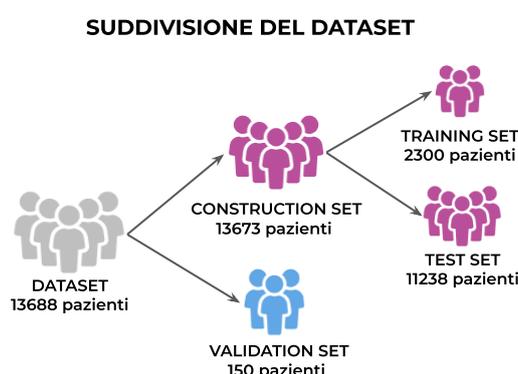
Da questa analisi, effettuata per ciascuna variabile, è stato notato che il valore del parametro k più piccolo, comporta una variazione inferiore nella distribuzione dei valori delle variabili a seguito del processo di imputazione, per entrambe le classi. È stato quindi scelto di mantenere per ogni variabile i valori sostituiti tramite l'impiego del parametro k pari a 9 e 11. Con particolare riferimento alla variabile presa come esempio nella figura 8 è possibile riportare la percentuale di valori mancanti pari al 52,2% della totalità dei pazienti, in dettaglio suddivisi come 970 pazienti appartenenti alla classe delle estubazioni fallite e 6172 pazienti appartenenti alla classe delle estubazioni riuscite. Al fine di eseguire il processo di sostituzione tramite il metodo del kNN è stato impiegato, nel caso della variabile presa come esempio, un training set di numerosità pari a 1769 pazienti per la classe delle estubazioni fallite e 4249 pazienti per la classe delle estubazioni riuscite. Dall'osservazione dei grafici nella figura 8 è anche possibile notare una distribuzione dei valori della concentrazione di albumina nel sangue più spostata verso valori maggiori per quanto riguarda la classe delle estubazioni riuscite rispetto ai pazienti che non hanno tollerato lo svezzamento.

In seguito alle operazioni di data cleaning effettuate è stato possibile ottenere il dataset definitivo, comprensivo di 85 variabili e completo in tutti i campi, implicato nelle fasi successive per la costruzione dei modelli di machine learning.

## Capitolo 5

### Suddivisione del dataset

La prosecuzione dello sviluppo del sistema di aiuto alla diagnosi vede coinvolta la fase di costruzione di un classificatore in grado di distinguere gli elementi in due classi, per questo particolare problema clinico. La costruzione del classificatore prevede l'utilizzo degli elementi del dataset per mettere in atto le fasi di allenamento, testing e validazione di un modello. A questo fine il dataset viene suddiviso in tre porzioni, ognuna delle quali verrà utilizzata per eseguire le tre fasi della costruzione del classificatore (figura 9).



*Figura 9. Rappresentazione della suddivisione del dataset in 3 sottoinsiemi di pazienti dedicati alla fase di apprendimento (training set), testing (test set) e validazione (validation set).*

In questa fase di suddivisione del dataset viene tenuto conto di alcune caratteristiche degli elementi presenti nel dataset, in particolare, della classe dell'estubazione e del fatto che alcuni pazienti hanno subito l'imputazione del valore di alcune variabili. Quindi l'intero dataset viene considerato come l'insieme di 4 sottogruppi: i pazienti di classe ES che hanno subito imputazione, i pazienti di classe ES che non avevano missing values, i pazienti di classe EF che hanno subito imputazione e i pazienti di classe EF che non avevano missing values.

Inizialmente, sono stati estratti dall'intero dataset 150 pazienti che verranno implicati solamente nella fase finale di validazione del modello. Questa estrazione è di tipo randomico ed in particolare sono stati estratti randomicamente 50 pazienti per ciascuno dei due sottogruppi della classe ES e 25 pazienti per ciascuno dei due sottogruppi della classe EF. In questo modo si può essere sicuri di tenere conto delle diverse caratteristiche, note a priori, che appaiono nei pazienti del dataset.

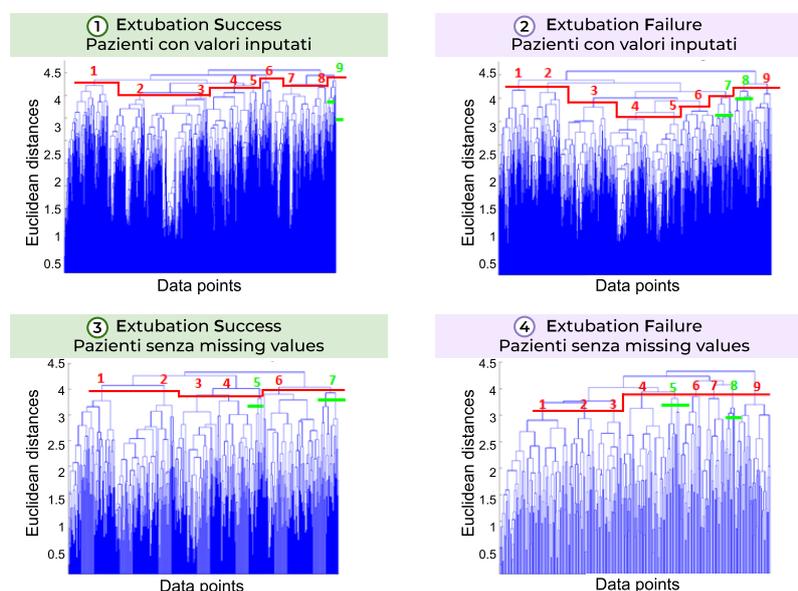
Tutti i restanti pazienti sono quelli che compongono il construction set, cioè quell'insieme di elementi implicati nella costruzione effettiva del modello. Questo viene a sua volta suddiviso in un training set e un test set utilizzati, rispettivamente, per la fase di allenamento e testing del classificatore. È importante l'estrazione degli elementi che andranno a comporre il training set, in quanto questo dovrà essere il più possibile rappresentativo dell'intero dataset e dovrà rappresentare in egual modo ciascuna delle due classi di pazienti. Per poter fare questo, i quattro sottogruppi individuati contenenti gli elementi del construction set sono stati sottoposti a un processo di clustering. Dunque

ogni sottogruppo è stato suddiviso in ulteriori sottoinsiemi con la volontà di avere dei cluster diversi tra di loro, ma contenenti ciascuno elementi simili tra loro. È stata utilizzata una tecnica di clustering gerarchico che costruisce un dendrogramma.

La costruzione del dendrogramma inizia considerando ogni elemento come se fosse un cluster a se stante e calcola le distanze (euclidee nello svolgimento di questo progetto di tesi) tra tutte le coppie possibili di elementi. A questo punto si uniscono gli elementi che presentano la distanza minore tra tutte quelle calcolate e viene quindi costituito un nuovo cluster. Ottenuti dei cluster formati da più di un elemento è necessario definire il prototipo di un cluster, che nel caso più semplice può essere il valore medio degli elementi al suo interno. Per calcolare poi una distanza tra un elemento e un cluster si sfrutterà il prototipo di quel cluster che sarà per esso rappresentativo. Nel presente lavoro per calcolare un valore di distanza tra due cluster si è calcolata la distanza euclidea tra gli elementi più lontani dei cluster. Si prosegue quindi, aggregando gli elementi in base al calcolo delle distanze, finché si avrà un solo gruppo in cui sono riuniti tutti gli elementi. Il dendrogramma tiene traccia di tutte le consecutive unioni e permette di visualizzare l'organizzazione degli elementi vedendo raggruppati quelli più simili.

È stato costruito un dendrogramma per ciascuno dei quattro sottogruppi considerati del construction set (figura 10). Al fine di eseguire l'analisi gerarchica, i valori di ogni variabile descrivente gli elementi del construction set, sono stati riportati in un range di valori compreso tra 0 e 1.

#### CLUSTERING GERARCHICO, TRAMITE TECNICA DEL DENDROGRAMMA, DEL CONSTRUCTION SET



*Figura 10. Rappresentazione dei dendrogrammi costruiti per ciascuno dei quattro sottogruppi individuati tra i pazienti del construction set tenendo conto di caratteristiche note a priori. È indicato in rosso, per ciascun dendrogramma, il taglio che permette di ottenere la suddivisione in clusters, mentre in verde è indicata l'ulteriore suddivisione di quei clusters che presentavano una variabilità elevata rispetto agli altri individuati.*

Analizzando questi diagrammi si tenta una prima operazione di taglio dei dendrogrammi volta a identificare dei sottogruppi distinti di pazienti simili, che è rappresentata nella figura con le linee orizzontali rosse che intersecano le linee di unione del dendrogramma. Questa prima suddivisione porta ad avere 9 cluster per il sottogruppo dei pazienti che presentano valori imputati sia per la classe ES che per la classe EF, 7 cluster per il sottogruppo dei pazienti di classe ES che non avevano valori mancanti e 9 cluster per il corrispettivo gruppo della classe EF. Questi sono stati valutati in termini di variabilità tra gli elementi costituenti un singolo cluster (variabilità intra-cluster), in quanto ci si pone l'obiettivo di applicare dei tagli che creino dei cluster con valori di variabilità intra-cluster comparabile (figura 11). A questo proposito i cluster che presentavano una variabilità elevata rispetto agli altri sono stati ulteriormente suddivisi andando a spostare più in basso nel diagramma il taglio, fino ad ottenere dei cluster con variabilità paragonabile a quella degli altri cluster.

#### INDIVIDUAZIONE DEI CLUSTERS DAI DENDROGRAMMI

		CLUSTERS								
		1	2	3	4	5	6	7	8	9
①	Variabilità intra-cluster	1.85	1.53	1.51	1.70	1.86	1.86	1.75	1.83	1.99
	Numerosità del cluster	1618	1367	2321	1613	292	529	1432	660	321
②	Variabilità intra-cluster	1.85	1.92	1.69	1.58	1.47	1.84	1.99	2.02	1.93
	Numerosità del cluster	273	286	422	401	263	324	149	198	171
③	Variabilità intra-cluster	1.87	1.72	1.75	1.85	2.04	1.85	2.04		
	Numerosità del cluster	100	128	91	81	17	125	38		
④	Variabilità intra-cluster	1.72	1.79	1.75	1.74	2.05	1.89	1.90	2.01	1.93
	Numerosità del cluster	80	31	32	53	16	24	28	12	42

Cluster 9									
1.77	1.81	1.86	1.80	1.86	1.66	1.81			
5	9	129	137	21	11	9			

Cluster 7				Cluster 8		
1.76	1.79	1.80	1.85	1.69	1.90	1.88
9	31	51	58	13	67	118

Cluster 5		Cluster 7	
1.72	1.81	1.84	1.92
7	10	19	19

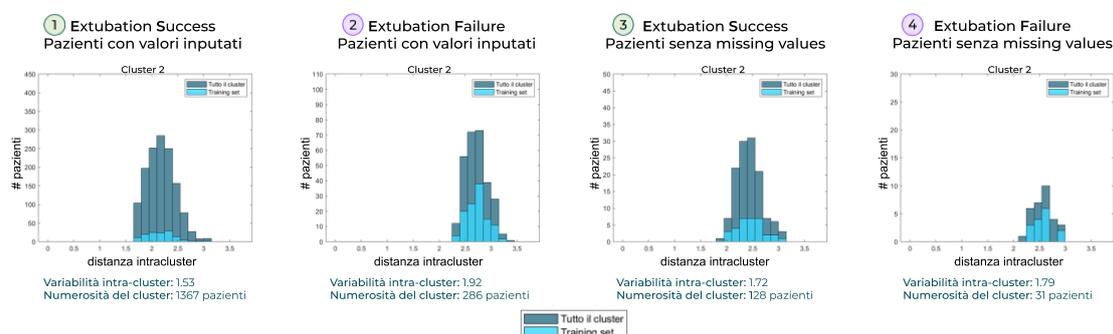
Cluster 5			Cluster 8	
1.90	1.88	1.37	1.77	
11	5	2	9	1

**Figura 11.** Sono riportati nella figura i valori di variabilità tra gli elementi costituenti ogni singolo cluster e le numerosità di tali clusters. Nelle tabelle di sinistra sono presenti i dettagli dei clusters individuati dal primo tentativo di taglio dei dendrogrammi rappresentati in figura 10. In queste tabelle sono evidenziati in giallo i clusters che presentano una variabilità intra-cluster non paragonabile a quella di tutti gli altri clusters. Essi sono stati, allora, ulteriormente suddivisi fino a ottenere sottoinsiemi di elementi con omogeneità confrontabile a quella di tutti gli altri sottoinsiemi individuati (tabelle di destra). Infine, tutti i cluster evidenziati in verde sono quelli definitivamente ottenuti.

In questo modo è stato possibile ottenere dei gruppi omogenei di pazienti ed estraendo da ciascuno di questi alcuni pazienti è possibile ottenere un insieme di pazienti rappresentativi della totalità dei pazienti a disposizione. In particolare, è stato effettuato un campionamento randomico di un numero di elementi proporzionale alla dimensione del cluster, da ciascun cluster ricavato. Ad esempio, se voglio estrarre 1000 elementi dal totale degli elementi presenti in un sottogruppo e questo mostra un cluster con numerosità pari alla metà di tutti gli elementi di quel sottogruppo, estrarrò da tale cluster 500 elementi. In conclusione, sono stati estratti 1000 pazienti dal sottogruppo 1, 1000 pazienti dal sottogruppo 2, 150 dal sottogruppo 3 e 150 dal sottogruppo 4.

La rappresentatività degli elementi estratti da un cluster rispetto alla totalità del cluster è stata verificata rappresentando la distribuzione di tutti gli elementi del cluster, in termini di distanze tra questi elementi, confrontandola con la distribuzione di quelli da esso estratti.

RAPPRESENTATIVITÀ DEGLI ELEMENTI ESTRATTI PER COSTITUIRE IL TRAINING SET RISPETTO ALLA TOTALITÀ DEL CLUSTER  
ISTOGRAMMI DISTANZE INTRA-CLUSTER



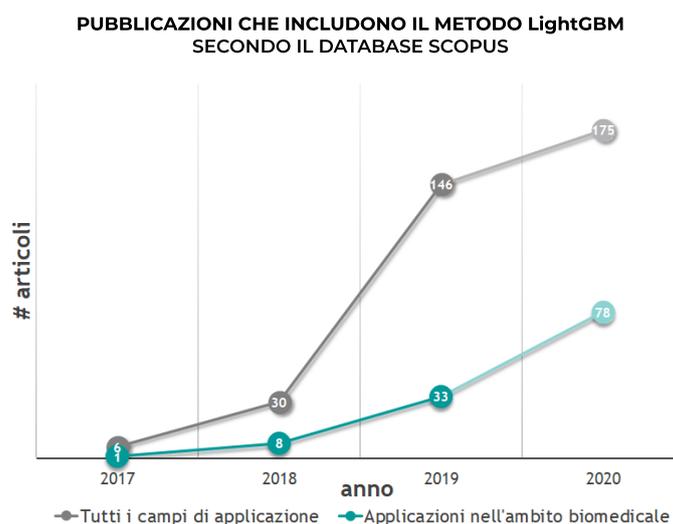
*Figura 12.* Sono riportati in figura, a titolo di esempio, alcuni grafici che mostrano il confronto della distribuzione di tutti i pazienti di un cluster e di solo quei pazienti estratti da esso per costituire il training set, caratterizzandoli secondo le mutue distanze tra loro. È mostrato l'esempio del secondo cluster individuato per ciascuno dei quattro sottogruppi del construction set.

Questa procedura ha permesso di creare un training set rappresentativo dell'intero dataset, contenendo al suo interno esempi di ogni tipologia di paziente presente nel dataset. Inoltre il training set ricavato presenta un egual numero di pazienti per ciascuna delle due classi, dunque è bilanciato tra le due classi e questo aiuterà nella fase di allenamento del classificatore a non ottenere un comportamento del modello sbilanciato tra le due classi. Infine, tutti gli elementi restanti hanno costituito il test set che verrà usato nella fase di testing del modello allenato.

## Capitolo 6

### Costruzione del classificatore Light Gradient Boosting Machine

Il 1° Ottobre 2016 la Microsoft ha introdotto una nuova tecnica di machine learning, chiamata Light Gradient Boosting Machine (LightGBM). Questa è una tecnica di apprendimento supervisionato, che mira quindi ad estrarre conoscenza da un insieme di dati noti usati per l'apprendimento automatico e ad applicare tale conoscenza su dati nuovi per scopi di classificazione o regressione. Il LightGBM ha subito un coinvolgimento sempre maggiore nei progetti di ricerca, in seguito alla sua introduzione nel mondo del machine learning. Si può notare un uso non indifferente di questa tecnica anche nelle applicazioni di carattere biomedicale, come la bioinformatica, la farmacologia, l'analisi di segnali elettromiografici ed elettroencefalografici e diversi problemi clinici (figura 13).



*Figura 13.* Nella figura è mostrato il numero di pubblicazioni che impiegano la tecnica di machine learning 'Light Gradient Boosting Machine' nei loro studi, nel corso degli anni fino al 2020. È possibile notare tra la totalità delle pubblicazioni (curva grigia) quelle che trattano problemi nell'ambito biomedicale (curva verde).

Più in dettaglio si tratta di *ensemble learning*, che è un metodo di apprendimento che costruisce un modello (*strong learner*) tramite la composizione di vari modelli più semplici (*weak learners*). Uno *strong learner* è un modello che tenta di avvicinarsi alle prestazioni ideali, mentre un *weak learner* è un modello che ha prestazioni leggermente superiori a quelle della casualità. L'*ensemble learning* si fonda quindi sull'idea che la cooperazione tra diversi modelli deboli riesca a massimizzare le prestazioni complessive, sfruttando i punti di forza e limitando le debolezze dei singoli modelli, risultando quindi nel comportamento come un unico modello forte. In particolare, nel metodo LightGBM il *weak learner* è rappresentato da un albero decisionale e l'algorithm iterativo tramite il quale si costruisce lo *strong learner* è rappresentato dal *boosting*.

Applicando questa tecnica ad ogni iterazione viene costruito un weak learner che viene aggiunto a quelli già presenti per comporre il modello finale. Questo approccio è detto *stagewise*.

$$\mathcal{L}_n(x) = \mathcal{L}_{n-1}(x) + \ell_n(x)$$

$\ell_n(x)$  è un generico weak learner

$\mathcal{L}_{n-1}(x)$  è il modello costruito nell'iterazione precedente

$\mathcal{L}_n(x)$  è il modello costruito allo step  $n$

Nella specifica tecnica del LightGBM i decision trees vengono allenati per correggere gli errori di predizione fatti dal modello all'iterazione precedente e così ogni nuovo weak learner punta a stimare i residui che, via via, si riducono progressivamente .

Inoltre, come indica il nome, il LightGBM implica l'uso del *gradient boosting*, che ha lo scopo di minimizzare una funzione di costo attraverso il meccanismo del gradient descent. A seconda del tipo di problema da risolvere si può usare una qualsiasi funzione di costo differenziabile, in quanto per il calcolo dei residui interviene la derivata parziale di tale funzione. Per eseguire la procedura gradient descent è necessario aggiungere ad ogni iterazione un decision tree al modello, in modo tale che riduca la funzione di costo.

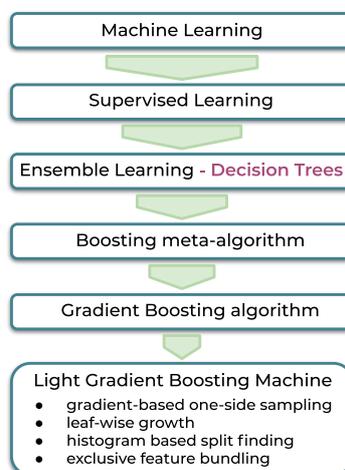
Un aspetto importante del gradient boosting è rappresentato dalle tecniche di regolarizzazione che mirano a ridurre il fenomeno dell'overfitting. Questo fenomeno si manifesta nell'allenamento del LightGBM e influenza quindi la capacità di generalizzazione del modello. Esistono, però, alcuni metodi che tentano di ridurre l'effetto dell'overfitting. Il primo, conosciuto come *incremental shrinkage*, consiste nell'aggiunta di una costante moltiplicativa  $\lambda$ , chiamata *learning rate*, al valore di uscita dei weak learners.

$$\mathcal{L}_n(x) = \mathcal{L}_{n-1}(x) + \lambda \ell_n(x) \text{ con } \lambda \in (0,1]$$

L'introduzione del learning rate migliora la capacità di generalizzazione del modello, ma comporta un maggior numero di iterazioni necessarie e quindi un maggiore tempo computazionale. Un altro metodo è rappresentato dallo *stochastic gradient boosting*, che invece migliora sia la generalizzazione sia la velocità dell'allenamento. Questo è un'implementazione del gradient boosting nella quale, a differenza della versione tradizionale, ogni volta che un decision tree viene addestrato gli viene sottoposto solamente un sottoinsieme degli elementi appartenenti al training set, estratto in modo random senza reinserimento. Terminato l'allenamento, l'albero decisionale imputerà l'intero training set, in modo tale da ottenere tutti i valori per l'iterazione successiva. Esistono anche delle varianti che sottocampionano le features prima di creare un nuovo albero oppure sottocampionano le features prima di un nuovo split nella costruzione di un singolo albero. Un'ulteriore strategia, impiegabile quando si vuole fare regressione, è il *penalized gradient boosting*. Infatti, gli alberi decisionali che effettuano regressione presentano dei valori ai nodi foglia, chiamati pesi, e questi possono essere regolarizzati tramite la norma  $L_1$  o  $L_2$ .

Nella figura 14 di seguito rappresentata possiamo osservare la collocazione della tecnica LightGBM nel mondo del machine learning con livelli di dettaglio progressivi.

## PERCORSO GERARCHICO DEL METODO LightGBM



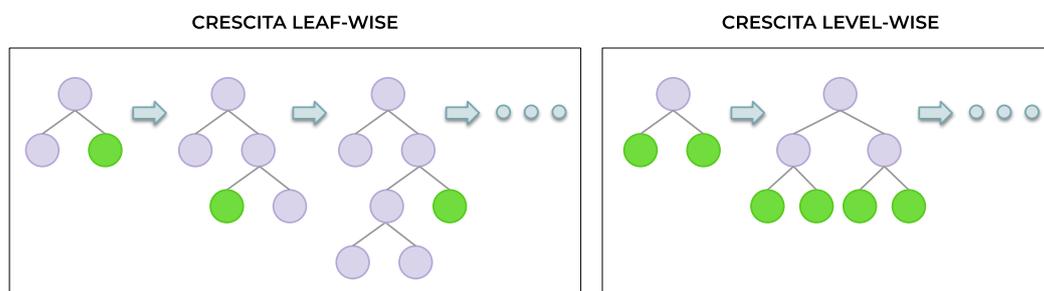
**Figura 14.** Classificazione del metodo Light Gradient Boosting Machine partendo dall'appartenenza all'insieme dei metodi di machine learning, passando per l'insieme dei metodi di apprendimento supervisionato, fino ad arrivare alla particolare categoria dei metodi di ensemble learning che sfruttano il principio del boosting.

Avendo terminato la descrizione delle categorie in cui rientra il LightGBM restano da spiegare le strategie tipiche di tale tecnica.

Il LightGBM sfrutta un metodo di campionamento degli elementi del training set basato sul modulo del gradiente, chiamato *gradient-based one-side sampling* (goss). Questo metodo prende tutti i campioni con un alto gradiente (alto training error) ed esegue un'estrazione random su quelli con un basso gradiente (basso training error). Permette di dare più importanza agli elementi difficili da imparare, garantendo il mantenimento della distribuzione originale dei dati. Questo metodo permette di convergere più velocemente, ma può andare incontro a overfitting se il dataset non è abbastanza numeroso.

Il LightGBM applica, inoltre, un algoritmo basato sull'istogramma per trovare lo split point ottimale nella creazione degli alberi. Ogni variabile numerica continua viene, infatti, splittata in bin discreti e la complessità computazionale risulta, quindi, paragonabile a **#campioni × #features**. Su questo aspetto entra in gioco la tecnica *exclusive feature bundling*, che identifica le variabili mutuamente esclusive (che hanno valori non nulli non sovrapposti) e le raggruppa in un'unica feature riducendo così la complessità.

Infine, il LightGBM si distingue da altri modelli basati sul gradient boosting, come il XGBoost, anche nella strategia impiegata per accrescere gli alberi decisionali. La costruzione degli alberi segue, infatti, una crescita *leaf-wise* (figura 15).



**Figura 15.** Rappresentazione grafica dei due metodi di accrescimento degli alberi decisionali. I nodi evidenziati in verde sono quelli individuati allo step corrente per proseguire l'accrescimento dell'albero decisionale. Nella tipologia di crescita level-wise ad ogni step si aggiunge un livello di profondità dell'albero, mentre nella crescita leaf-wise questo può non accadere nel caso in cui viene individuato come nodo con il più alto splitting gain uno appartenente a un livello precedentemente sviluppato.

Ad ogni step della crescita, viene identificato il nodo con il più alto splitting gain tra tutte le foglie e da qui viene accresciuto l'albero decisionale. Rispetto al metodo *level-wise* usato dalla tecnica XGBoost, questo permette di convergere più velocemente, ma le possibilità di andare incontro a overfitting aumentano. Infatti, a parità di numero di nodi terminali (foglie) un albero accresciuto con la strategia leaf-wise è più profondo di uno accresciuto con la strategia level-wise, che sostanzialmente crea un albero di forma piramidale.

È possibile ancora personalizzare la fase di crescita di un albero decisionale impostando il valore di alcuni parametri. Il primo è il numero massimo di nodi terminali che ogni decision tree può raggiungere e come effetto sull'allenamento si ha che a un maggior numero di foglie corrisponde un aumento dell'accuratezza sul training set, ma anche una maggiore possibilità di andare incontro a overfitting. Un altro parametro che ha influenza sul numero ottimale di quello appena descritto è la profondità massima di ogni albero; infatti è consigliato effettuare un tuning di questi due parametri insieme, tenendo anche conto che il metodo leafwise genera un albero più profondo del metodo levelwise. Anche in questo caso, una profondità elevata degli alberi porta a overfitting. Infine, è possibile indicare il numero minimo di elementi del training set che vanno a finire in una foglia, in modo tale da non andare a classificare un numero estremamente ridotto di elementi con quel ramo, che porterebbe nuovamente a una perdita di generalizzazione.

Ritornando a una visione complessiva dell'algoritmo iterativo implementato nella tecnica del LightGBM è possibile personalizzare alcuni aspetti della fase di allenamento (figura 16).

## DIAGRAMMA DI FUNZIONAMENTO DEL LightGBM

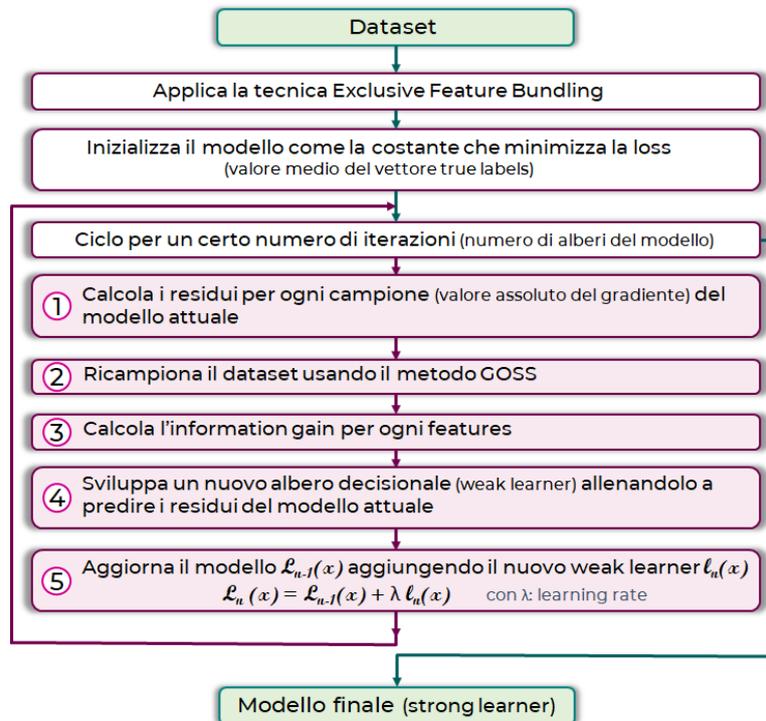


Figura 16. Diagramma di funzionamento dell'iterativo impiegato dalla tecnica LightGBM.

Si può impostare il numero di alberi da costruire, cioè il numero di iterazioni da effettuare siccome ad ogni iterazione viene costruito un nuovo albero. È consigliato usare valori bassi di learning rate con un alto numero di iterazioni. L'allenamento può essere arrestato nel caso in cui una metrica calcolata sul test set non migliora in un numero prestabilito di ultime iterazioni (*early\_stopping\_rounds*). È consigliato impostare tale parametro pari al 10% del numero di iterazioni, tenendo conto che un valore troppo alto incrementa il rischio di overfitting. Terminata la fase di allenamento, e dunque raggiunto il numero massimo di iterazioni impostato, viene scelto il modello costruito all'iterazione che restituisce il valore migliore della metrica calcolata sul test set. È possibile personalizzare anche quest'ultima. Alcune disponibili metriche sono: *MAE(11)*, *MSE(12)*, *RMSE*, *Huber loss*, *AUC*, *kullback\_leibler* o *binary\_logloss*. Altrimenti il parametro è impostabile come *null* se si vuole specificare una propria metrica tramite l'impiego di un ulteriore parametro (*feval*).

Un'ulteriore caratterizzazione riguarda le variabili categoriche, in quanto esse non sono gestite tramite una codifica one-hot, che porterebbe a dover avere un albero molto profondo per poter raggiungere una buona accuratezza. Viene usato, invece, un metodo che separa le categorie di una variabile categorica in 2 sottogruppi ed avendo  $k$  categorie si possono avere  $2^{k-1} - 1$  combinazioni. L'istogramma viene ordinato rispetto ai suoi valori accumulati e quindi ne trova lo split ottimale applicando il metodo di Fisher

(sviluppato nel 1958). È consigliato indicare manualmente quali sono le features categoriche, inserendo gli indici delle colonne.

L'algoritmo permette, inoltre, di gestire il bilanciamento delle prestazioni del modello rispetto le classi. Settando il parametro *is\_unbalance* a *True* l'algoritmo prova automaticamente a bilanciare il peso della classe dominante oppure a questo fine si può impostare il parametro *scale\_pos\_weight* come il rapporto tra il numero di campioni della classe negativa e quello della classe positiva.

Il LightGBM può svolgere due compiti di apprendimento, regressione o classificazione. Proprio quest'ultima modalità è stata impiegata per l'estrazione di conoscenza dai dati relativi agli eventi di ventilazione meccanica invasiva, permettendo di classificarli in due categorie in base all'esito dell'estubazione.

## 6.1 Apprendimento e validazione

Nel presente progetto di tesi è stata effettuata la costruzione del modello di machine learning Light Gradient Boosting Machine tramite l'utilizzo della libreria *lightgbm* rivolta al linguaggio Python. Nella sezione precedente è stata mostrata una panoramica sui parametri caratteristici del modello di machine learning LightGBM. Questi parametri devono essere impostati in modo tale da adattare il modello ai particolari dati che deve analizzare per cercare di raggiungere le migliori prestazioni possibili. A questo scopo è indispensabile un processo di tuning degli iperparametri del modello, che li ottimizza rispetto al problema che si vuole trattare. In particolare, alcuni sono direttamente determinabili a seguito della comprensione del problema da trattare o naturalmente stabiliti dall'obiettivo di apprendimento che ci si pone, come l'obiettivo di classificazione binaria e l'impostazione del metodo di boosting *goss* che permette di eseguire l'implementazione del LightGBM anziché altri algoritmi di boosting o ensemble learning. I due parametri tipici del metodo *goss*, *top\_rate* e *other\_rate*, sono stati scelti tenendo conto che la somma tra i due valori deve essere non superiore all'unità e deve permettere di selezionare non pochi elementi dall'insieme di quelli dedicati al training. Con lo stesso fine, che mira a ridurre il fenomeno dell'overfitting che risulta ampliato nella tecnica del LightGBM se il numero di esempi su cui viene allenato il modello non è elevato, è stato impostato il parametro *subsample* pari a 1, così da fornire l'intero training set agli alberi decisionali durante il loro addestramento. Sono di seguito elencati i parametri fissati, che non verranno sottoposti al processo di tuning:

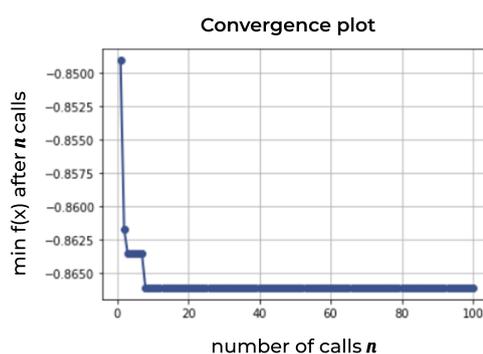
- objective: *'binary'*
- metric: *'auc'*
- is\_unbalance: *True*
- boosting: *'goss'*
- num\_trees: 500
- early\_stopping\_rounds: 50
- subsample: 1
- top\_rate: 0.6
- other\_rate: 0.3

Altri 5 iperparametri sono stati ottimizzati congiuntamente effettuando un'esplorazione nello spazio di ricerca definito dagli intervalli di valori assumibili da questi iperparametri. Gli intervalli tra cui sono stati cercati i valori ottimali sono i seguenti:

- learning rate: [ 0.01 , 0.1 ]
- max\_depth: [ 1 , 30 ]
- num\_leaves: [ 10 , 200 ]
- min\_data\_in\_leaf: [ 10 , 50 ]
- feature fraction: [ 0.1 , 1 ]

Nell'esperimento sono state provate 100 diverse configurazioni dei parametri a partire da 10 set di parametri scelti a caso. Ogni prova consiste nell'allenamento del modello sul training set e valutazione delle sue prestazioni sul test set con un determinato set di parametri. Il tuning ha lo scopo di minimizzare una funzione obiettivo, definita in questo caso come  $f(x) = -1 \times \text{AUC sul test set}$ . Il processo di ottimizzazione sequenziale è stato eseguito sfruttando la libreria Scikit-Optimize. È possibile osservare la curva di convergenza di tale processo nel grafico sottostante, che mostra l'andamento del minimo momentaneamente trovato della funzione obiettivo nel corso delle 100 prove (figura 17).

**GRAFICO DI CONVERGENZA DEL PROCESSO DI TUNING DEGLI IPERPARAMETRI**



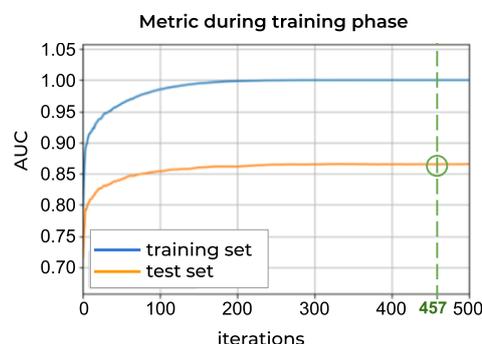
*Figura 17. Curva di convergenza del processo di ottimizzazione degli iperparametri del modello LighGBM.*

Al termine del processo di tuning sono stati selezionati i seguenti valori degli iperparametri cercati:

- learning rate: 0.032
- max\_depth: 14
- num\_leaves: 131
- min\_data\_in\_leaf: 48
- feature\_fraction: 0.39

A questo punto è stata eseguita la fase di allenamento del modello, specificando tutti i parametri definiti. L'avanzamento dell'allenamento è stato monitorato valutando la metrica, calcolata sia sul training set che sul test set, nel corso delle iterazioni (figura 18).

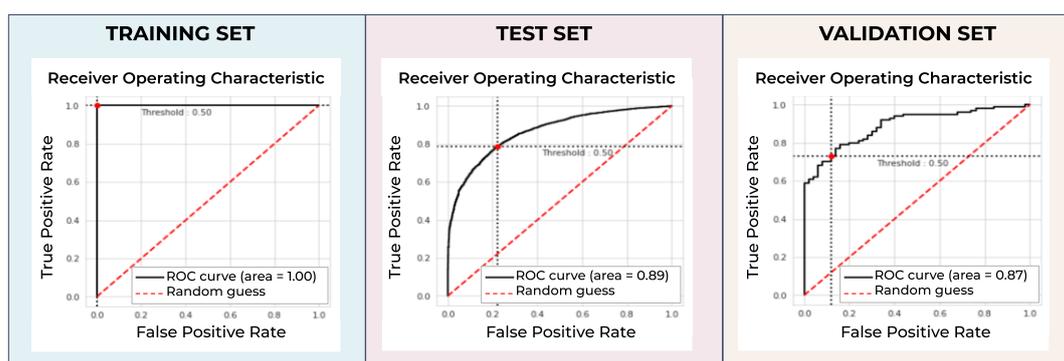
## METRICA DURANTE IL PROCESSO DI APPRENDIMENTO DEL MODELLO



**Figura 18.** È rappresentata in figura la metrica (Area Under the ROC Curve) calcolata sia sull'insieme dei pazienti appartenenti al training set (curva blu), sia sull'insieme dei pazienti appartenenti al test set (curva arancione), in ciascuna iterazione fino al numero massimo di iterazioni impostato. È possibile, così, osservare il comportamento del modello durante la fase di apprendimento. Il modello costruito all'iterazione 457 è stato scelto come modello finale in quanto nelle iterazioni successive non si ottengono più miglioramenti della metrica calcolata sul test set.

È stato calcolato, quindi, ad ogni iterazione il valore dell'area sottesa dalla curva ROC (AUC) del modello costruito nell'iterazione corrente. Il numero di iterazioni è stato impostato in modo tale da permettere la creazione di un modello sufficientemente complesso. Avendo la volontà di continuare ad ottenere dei miglioramenti di prestazioni sul test set, è stato impostato il parametro che permette di arrestare l'allenamento nel caso in cui la curva relativa al test set raggiunge un plateau. Al termine delle 500 iterazioni è risultato che il modello che raggiunge il valore migliore di AUC valutata sul test set è quello costituito da 457 alberi decisionali. Dunque tale modello è riconosciuto come modello finale al termine dell'allenamento.

A seguito della costruzione del classificatore è importante valutare le sue prestazioni sull'insieme dei dati usati per l'apprendimento e quelli usati per il testing ed, inoltre, eseguire una validazione del modello costruito verificandone le prestazioni su pazienti nuovi. Innanzitutto, sono state rappresentate le curve ROC ottenute testando il modello sui tre set di dati (figura 19).



**Figura 19.** In figura sono rappresentate le curve ROC costruite tramite le predizioni del modello rispetto ai pazienti ad esso sottoposti per ciascuno dei tre sottoinsiemi (training set, test set, validation set). È possibile notare il valore della area sottesa a ciascuna curva, in particolare pari a 1 per quanto riguarda il training set, pari a 0.89 per quanto riguarda il test set e pari a 0.87 per quanto riguarda il validation set.

Osservando le curve ROC è possibile notare una forte differenza tra le prestazioni sul training set e quelle sugli altri due insiemi di pazienti, che era già possibile notare dalla figura tot. L'area sottesa dalla curva ROC, costruita valutando il training set, raggiunge addirittura un valore pari a 1 e ciò significa che il modello si adatta perfettamente ai campioni presenti in tale dataset, ma non riesce ovviamente a mantenere tali prestazioni per pazienti su cui non si è allenato.

L'output del classificatore costruito è per ogni paziente valutato compreso tra 0 e 1, quindi fornisce la probabilità che il paziente sia appartenente alla classe positiva, dunque potrà tollerare l'estubazione. Al fine di ottenere una classificazione netta nelle due classi è stata identificata una soglia al di sopra della quale il paziente è riconosciuto come appartenente alla classe positiva. Considerando la curva ROC calcolata sulle predizioni del test set è stata scelta la soglia che permette di ottenere la maggiore differenza tra il tasso di veri positivi e quello di falsi positivi, cercando di collocarci, quindi, nel punto più in alto e a sinistra della curva ROC, che permette di godere contemporaneamente della maggiore sensibilità e specificità. Tale soglia, ottimizzata sugli elementi del test set, è pari a 0.506 ed è stata applicata, per effettuare la classificazione binaria, anche ai pazienti del training set e del validation set. Questo ha permesso di costruire per ciascun sottoinsieme di pazienti una confusion matrix (figura 20). Queste matrici indicano la percentuale di corretti classificati e percentuale di misclassificati rispetto al totale degli elementi della classe a cui appartengono. Tale rappresentazione, anziché i valori assoluti, è stata preferita in quanto il numero di elementi realmente appartenenti alle due classi è differente. Dunque, questa rappresentazione percentuale facilita il confronto e l'analisi delle prestazioni del modello.

TRAINING SET			TEST SET			VALIDATION SET			
Confusion matrix			Confusion matrix			Confusion matrix			
Thr=0.506			Thr=0.506			Thr=0.506			
			CLASSE REALE			CLASSE REALE			
CLASSE PREDETTA				EF	ES		EF	ES	
EF			100%	0%		78.4%	21.6%	88%	28%
ES			0%	100%		21.6%	78.4%	12%	72%

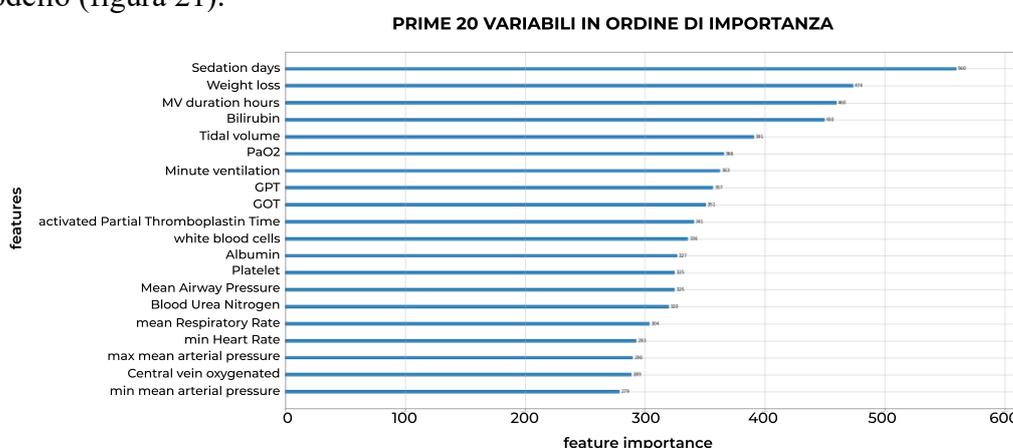
Figura 20. In figura sono riportate le confusion matrices ottenute dalla classificazione dei pazienti appartenenti al training set, al test set e al validation set.

È evidente il comportamento perfettamente bilanciato del modello rispetto alle due classi dei pazienti appartenenti al test set. Questo è dovuto all'attivazione di un approccio (isunbalance = 'True') del modello durante la fase di apprendimento, che cerca di bilanciare automaticamente il peso della classe dominante. Nonostante il training set sia stato costruito appositamente con un egual numero di pazienti che sono andati incontro al fallimento dell'estubazione e pazienti che avevano tollerato la respirazione spontanea, il modello ha la tendenza a riconoscere meglio i pazienti di classe negativa. Questo è stato riscontrato in un esperimento effettuato senza il bilanciamento automatico nell'allenamento del LightGBM e tale tendenza si può anche notare nelle prestazioni del modello sui pazienti dedicati alla validazione. Proprio in questa fase si può notare che il

modello mantiene pressoché la stessa capacità di corretta classificazione totale, rispetto alla fase di testing. Le prestazioni complessive del modello per pazienti su cui non si è allenato, sebbene si manifesti il fenomeno dell'overfitting, sono comunque accettabili.

## 6.2 Interpretazione del modello

Il modello LightGBM permette di effettuare un'interpretazione del processo di decisione che segue per restituire determinati risultati. Essendo un modello che si basa sulla costruzione di molteplici alberi decisionali è possibile, innanzitutto, valutare l'importanza delle variabili del dataset, a livello globale, calcolando il numero di volte che una variabile è utilizzata per effettuare uno split nella costruzione di tutti gli alberi che costituiscono il modello (figura 21).



*Figura 21.* Sono riportate in figura le prime 20 variabili ordinate secondo l'importanza globale calcolata come il numero di volte che la variabile viene usata per effettuare uno split nella costruzione degli alberi costituenti il modello. Una maggiore lunghezza della barra associata a una variabili indica un valore di importanza maggiore.

L'analisi di feature importance può fornire ulteriori dettagli sfruttando la tecnica SHAP (SHapley Additive exPlanations), che è in grado di mostrare quali features sono più influenti per la predizione, quali sono più rilevanti per la classe target e come le features interagiscono per raggiungere un alto livello informativo. Il valore di Shapley permette di assegnare l'importanza alle variabili in base al loro contributo nella valutazione totale che porta alla classificazione di un paziente in una determinata classe. È un metodo molto utile per interpretare la predizione di un singolo elemento. Ne è riportato un esempio nella figura 22 relativa al primo elemento del validation set che appartiene alla classe delle estubazioni fallite. Il colore blu indica un contributo negativo, mentre il colore rosso indica un contributo positivo per portare l'esito della predizione verso la classe Extubation Failure. Questo comportamento è intuibile anche dal verso in opposizione delle frecce nella sezione rossa e blu. La porzione, riguardante il contributo di ogni variabile, nella barra assume visivamente un'ampiezza maggiore per quelle variabili che hanno maggiore effetto nella predizione per il paziente considerato. In particolare, si può notare come la variabile che indica la durata della ventilazione meccanica invasiva ha il maggior contributo positivo sulla predizione. Il base value (-0.09427) corrisponde alla media dei valori di tutte le foglie del modello, cioè la predizione del modello nel caso in

cui non si considerasse nessuna feature, mentre il valore di output rispettivo della predizione del paziente in questione è pari a 2,97. Tale paziente è stato quindi correttamente classificato come appartenente alla classe delle estubazioni fallite.

**INTERPRETAZIONE DELLA PREDIZIONE DEL MODELLO PER UN SINGOLO ELEMENTO**



Figura 22. In figura è riportata l'interpretazione secondo Shapley della predizione del primo paziente appartenente al validation set.

È possibile anche creare un grafico che rappresenta il comportamento complessivo rispetto a tutti gli elementi di un dataset. Nella figura 23 ogni punto rappresenta un elemento del validation set. Sono visibili le prime 20 features rappresentate in ordine di importanza definita come il contributo che la variabile fornisce per effettuare le predizioni. È raffigurato anche in questo caso l'impatto che ciascuna features ha sulla classe target (EF anche in tale rappresentazione). Il colore varia da blu, che è associato ai valori bassi di una variabile nel suo range di appartenenza, a rosso che è associato, invece, ai valori alti. Sull'asse delle ascisse è indicato il valore di SHAP ed in questa rappresentazione a un valore positivo di SHAP corrisponde una relazione positiva con la classe delle estubazioni fallite.

**IMPORTANZA DELLE VARIABILI SECONDO SHAPLEY**

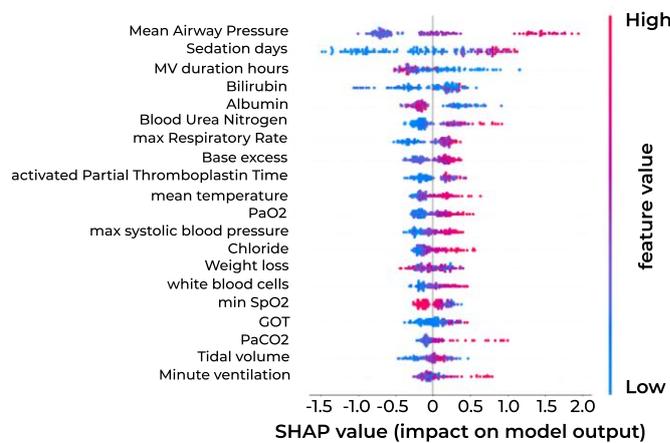


Figura 23. Raffigurazione del valore di SHAP e dunque dell'impatto delle variabili sulla predizione del modello considerando tutti i pazienti appartenenti al validation set. Sono rappresentate le prime 20 variabili ordinate secondo l'importanza definita dal metodo di Shapley.

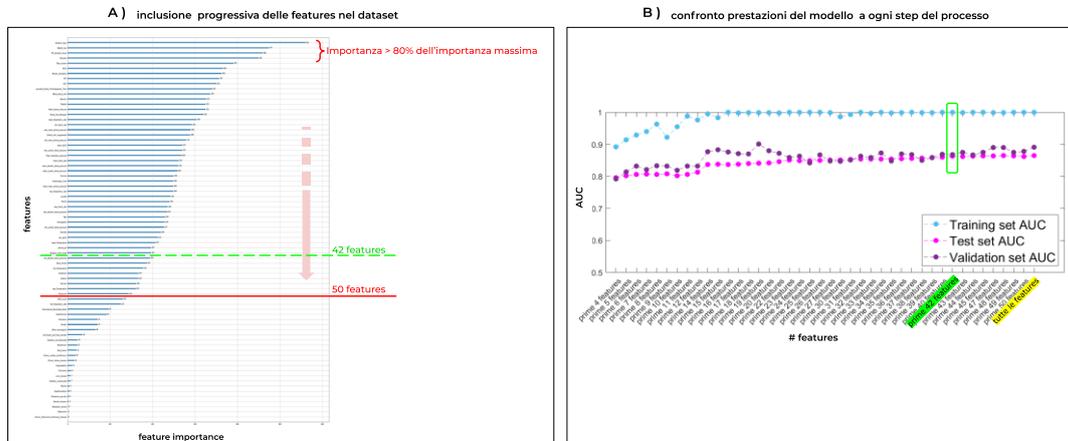
Osservando la figura 23 è possibile notare, ad esempio, che quando la variabile *Sedation\_days* assume valori elevati (rosso) ha un contributo positivo sulla predizione della classe *ExtubationFailure*, dunque numerosi giorni di somministrazione di farmaci sedativi portano la predizione del modello verso la classe delle estubazioni fallite. Secondo Shapley, allora, l'importanza di una variabile è proporzionale a quanto la predizione del modello varia considerando tale variabile, rispetto al caso in cui essa non venisse impiegata per descrivere i pazienti (in questo studio).

Il significato clinico delle variabili risultate importanti dalle appena descritte analisi, di importanza globale e importanza secondo Shapley, è stato confermato da un medico esperto, infatti tali variabili sono principalmente considerate durante la valutazione eseguita dagli esperti per stimare la probabilità del paziente di superare la prova di respirazione spontanea (SBT) e dunque tollerare lo svezzamento.

### 6.3 Forward Feature Selection

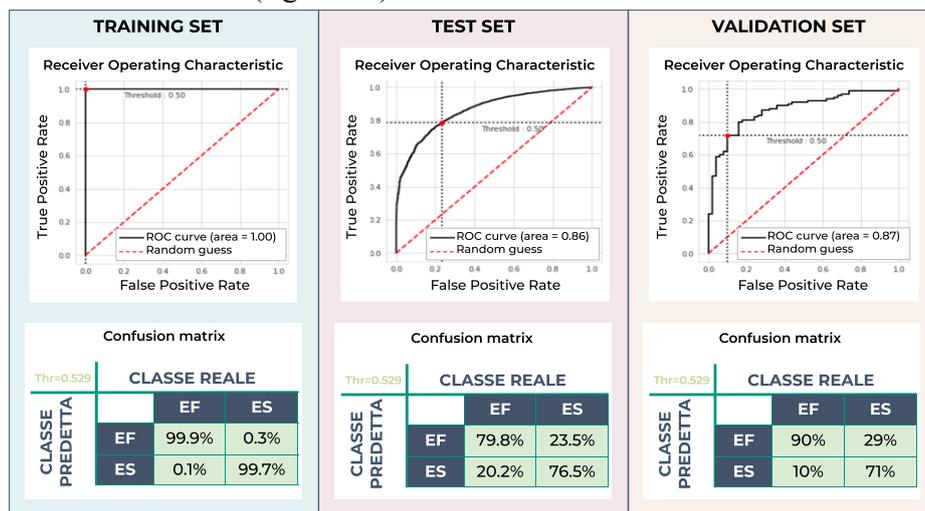
L'informazione sull'importanza delle variabili può essere sfruttata per progettare una procedura di feature selection, e dunque selezionare un sottoinsieme di variabili da quelle di partenza di numerosità minore possibile che sia in grado però di mantenere le prestazioni del classificatore allenato su tutte le variabili disponibili. In questo modo è possibile ridurre la dimensionalità del dataset originale e dunque la complessità del modello, la memoria occupata e i tempi di training del classificatore, ma a differenza di una feature construction il significato delle variabili è preservato e rimane possibile interpretare la conoscenza estratta dal modello. È stata implementata una strategia di ricerca del sottoinsieme di variabili, conosciuta in generale come forward feature selection, basata sull'importanza globale delle features precedentemente descritta (figura 24). In particolare, il punto di partenza della ricerca è stato individuato selezionando le variabili che presentano un'importanza superiore all'80% dell'importanza massima riscontrata tra le features. Esse sono 4 variabili: Sedation days, Weight loss, Mechanical ventilation duration hours e Bilirubin. Viene quindi costruito un classificatore sfruttando solamente tali variabili. Ad ogni step della procedura viene aggiunta una nuova variabile seguendo l'ordine determinato dall'importanza, o più di una variabile contemporaneamente nel caso in cui siano presenti più variabili con la stessa importanza, e viene allenato un modello con il subset di variabili considerato allo step corrente. Questa operazione è stata effettuata fino a raggiungere un sottoinsieme valutato di 50 variabili. È importante evidenziare che ad ogni step della ricerca è stato effettuando un tuning degli iperparametri del modello prima di intraprenderne l'allenamento, in modo tale che essi fossero ottimizzati rispetto al dataset su cui avveniva l'apprendimento. Tutti i modelli costruiti sono stati valutati tramite il calcolo di una metrica, l'area sottesa dalla curva ROC, valutata sul training set, sul test set e sul validation set. È stata raffigurata tale metrica rispetto l'aggiunta progressiva delle variabili confrontata anche, infine, con il modello costruito con tutte le 85 variabili disponibili.

ANDAMENTO DELLA METRICA DURANTE LA FORWARD FEATURE SELECTION BASATA SULL'IMPORTANZA GLOBALE DELLE VARIABILI



**Figura 24.** **A)** Ordinando tutte le variabili impiegate nel modello secondo l'importanza globale è possibile seguire tale ordinamento per attuare un processo di forward feature selection. In rosso si può osservare il sottoinsieme di variabili selezionato al primo step del processo e quello selezionato all'ultimo step dell'analisi costituito dalle prime 50 variabili. In verde si può notare il sottoinsieme scelto, composto da 42 variabili, e l'indicazione visiva della loro importanza (una maggiore estensione della barra blu corrisponde a una maggiore importanza). **B)** Confronto tra le prestazioni del modello costruito ad ogni step del processo di feature selection, calcolate sul training set, sul test set e sul validation set. In giallo è evidenziato il modello costruito impiegando tutte le 85 variabili, mentre in verde è evidenziato il modello scelto, costruito impiegando solamente 42 delle 85 variabili.

Ponendo attenzione soprattutto ai valori riferiti al test set, si può notare che già con il sottoinsieme di 42 features si ottengono prestazioni paragonabili a quelle ottenute descrivendo i pazienti con 85 variabili. Selezionando, allora, tale sottoinsieme di variabili è stato allenato un modello e ne sono state valutate le prestazioni, costruendo le curve ROC e le confusion matrices relative alla predizione degli elementi del training set, del test set e del validation set (figura 25).



**Figura 25.** In figura sono rappresentate le curve ROC costruite tramite le predizioni del modello rispetto ai pazienti ad esso sottoposti per ciascuno dei tre sottoinsiemi (training set, test set, validation set). È possibile notare il valore della area sottesa a ciascuna curva, in particolare pari a 1 per quanto riguarda il training set, pari a 0.86 per quanto riguarda il test set e pari a 0.87 per quanto riguarda il validation set. Sono, inoltre, riportate le confusion matrices ottenute dalla classificazione dei pazienti appartenenti al training set, al test set e al validation set.

## Capitolo 7

### Costruzione del classificatore “rete neurale”

Le reti neurali artificiali sono oggi ampiamente utilizzate per risolvere problemi legati a diversi ambiti, tra i quali anche il biomedicale nel quale si sono diffuse per svariate applicazioni.

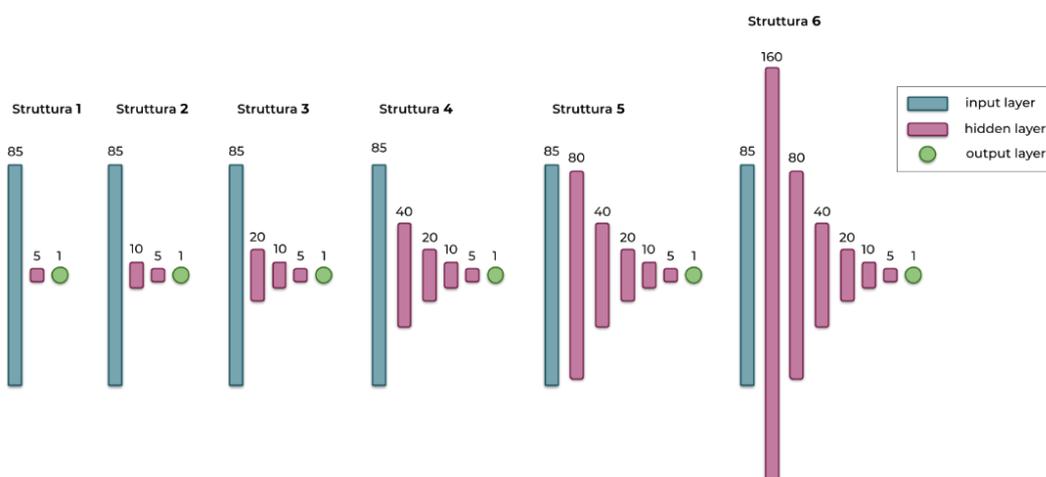
Nel presente progetto di tesi è stato scelto di applicare gli algoritmi delle reti neurali al problema dello svezzamento di un paziente dal ventilatore meccanico, strategia già comunemente sviluppata negli studi presenti in letteratura per affrontare tale problema clinico. In particolare, considerando la numerosità dei pazienti e delle variabili che li descrivono costituenti il dataset da analizzare, è stato intrapreso uno studio incentrato sulla tipologia di rete neurale definita Multi-Layers Perceptron.

Come la tecnica del LightGBM, trattata nel capitolo precedente, anche il Multi-Layers Perceptron è un modello di machine learning che permette di attuare un processo di apprendimento supervisionato. Tale rete neurale, a differenza del Perceptron, è adatta a riconoscere pattern complessi e a distinguere, quindi, classi di elementi non linearmente separabili. I suoi neuroni sono organizzati secondo un'architettura che prevede: 1 layer di input, 1 o più layers nascosti e 1 layer di output. Il layer di input è costituito da un numero di neuroni pari al numero delle variabili, mentre il layer di output è costituito generalmente da un solo neurone per i problemi di classificazione binaria. I layers nascosti possono avere un numero di neuroni variabile a seconda dell'applicazione, così come lo stesso numero di layers nascosti infrapposti tra quello di input e di output. Il Multi-Layers Perceptron è costituito da layers completamente connessi, cioè ogni neurone (nodo) in un layer è connesso, tramite un certo peso (arco), a ogni altro neurone del layer successivo. Ogni neurone è caratterizzato da una funzione di attivazione che viene applicata alla somma pesata degli input che riceve per produrre un output. L'algoritmo di apprendimento adoperato è la backpropagation, che mira a ottimizzare i pesi dei neuroni per raggiungere un errore dell'output restituito dalla rete rispetto alla classe reale, via via, inferiore rendendo il modello progressivamente più accurato.

#### 7.1 Apprendimento e validazione

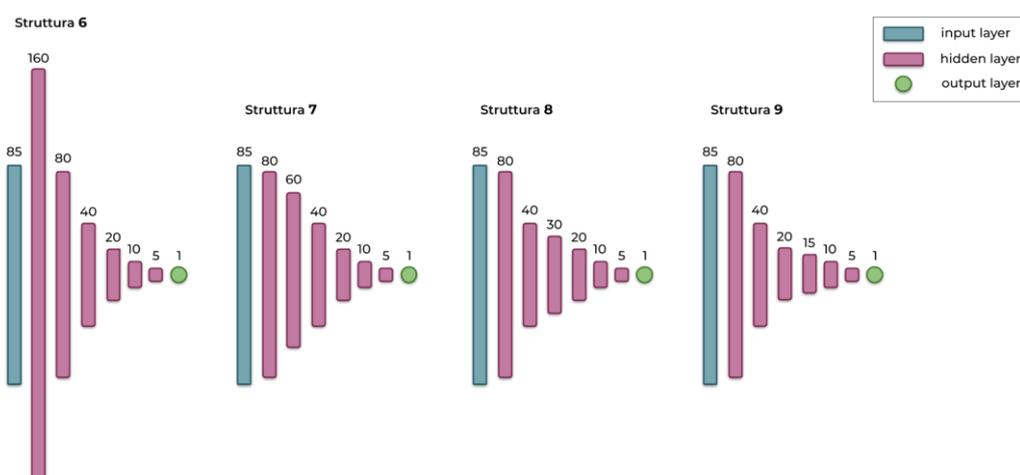
Nel presente studio è stata sfruttata l'implementazione delle feedforward neural networks fornita dal software Matlab. Questa prevede l'impiego, come funzione di attivazione per tutti i neuroni di ogni layer nascosto, della funzione sigmoidea tangente iperbolica (tansig) e per il neurone di output una funzione lineare (purelin).

Sono state analizzate diverse architetture gerarchiche, triangolari, di rete neurale ispirandosi alle consuetudini del deep learning. Per topologia triangolare si intende una struttura nella quale ogni layer nascosto possiede un numero di neuroni pari alla metà del precedente layer nascosto. In particolare, partendo da una struttura composta da un solo layer nascosto di 5 neuroni, sono state indagate le topologie triangolari con un numero di hidden layers crescente, fino a raggiungerne 6 (figura 26).



**Figura 26.** Rappresentazione delle topologie di rete neurale analizzate che presentano una struttura triangolare e possiedono rispettivamente un numero di hidden layers crescente. In blu è raffigurato, per ciascuna topologia, il layer di input costituito sempre da 85 neuroni, mentre in verde il layer di output costituito sempre da un solo neurone. In rosso sono raffigurati i layers nascosti ed in particolare la prima struttura ne possiede uno, la seconda due, e così via, fino a alla sesta struttura, aggiungendo ogni volta un layer che possiede un numero di neuroni pari al doppio del numero di neuroni dell'ultimo layer inserito.

Siccome il dataset studiato ha un'elevata numerosità di pazienti e variabili si è ritenuto che si potesse adattare meglio a una rete neurale con più layers nascosti e molti neuroni. Quindi, sono state indagate anche altre architetture composte da 6 layers nascosti, ma con una struttura non perfettamente triangolare (figura 27).



**Figura 27.** Rappresentazione delle topologie analizzate che possiedono 6 layers nascosti. In blu è raffigurato, per ciascuna topologia, il layer di input costituito sempre da 85 neuroni, in verde il layer di output costituito sempre da un solo neurone e in rosso sono raffigurati i layers nascosti.

Per ciascuna architettura sono state allenate 100 reti neurali, in modo tale da poter valutare le prestazioni in termini di valori medi percentuali ed errori standard percentuali. Questo è utile per analizzare un comportamento medio dell'architettura di rete e quindi valutarne la stabilità (figura 28). Per ciascuna rete neurale è stata costruita la curva ROC relativa al test set, sulla base della quale scegliere la soglia ottimale per effettuare la classificazione

binaria sia delle predizioni del test set, ma anche del training set. Tale soglia è quella che permette di ottenere il miglior compromesso tra sensibilità e specificità del modello. È importante porre in evidenza che per intraprendere la fase di apprendimento i valori di ogni variabile sono stati normalizzati in un intervallo di valori compresi tra 0 e 1, tramite la tecnica del min-max scaling, in quanto le variabili del dataset mostravano range di valori molto diversi tra loro. Le prestazioni sono state confrontate valutando, per ogni rete neurale costruita, la percentuale di corretta classificazione e l'errore percentuale di misclassificazione sia per la classe positiva, estubazioni riuscite, che per la classe negativa, estubazioni fallite.

CONFRONTO TRA LE PRESTAZIONI DELLE 9 ARCHITETTURE DI RETE NEURALE

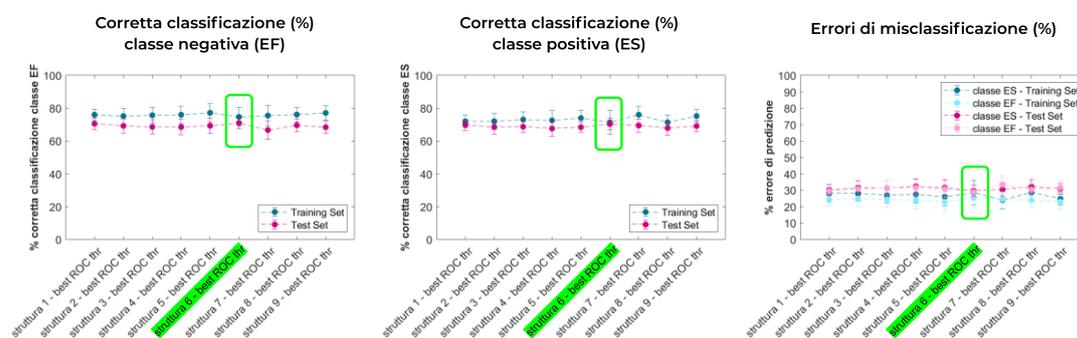
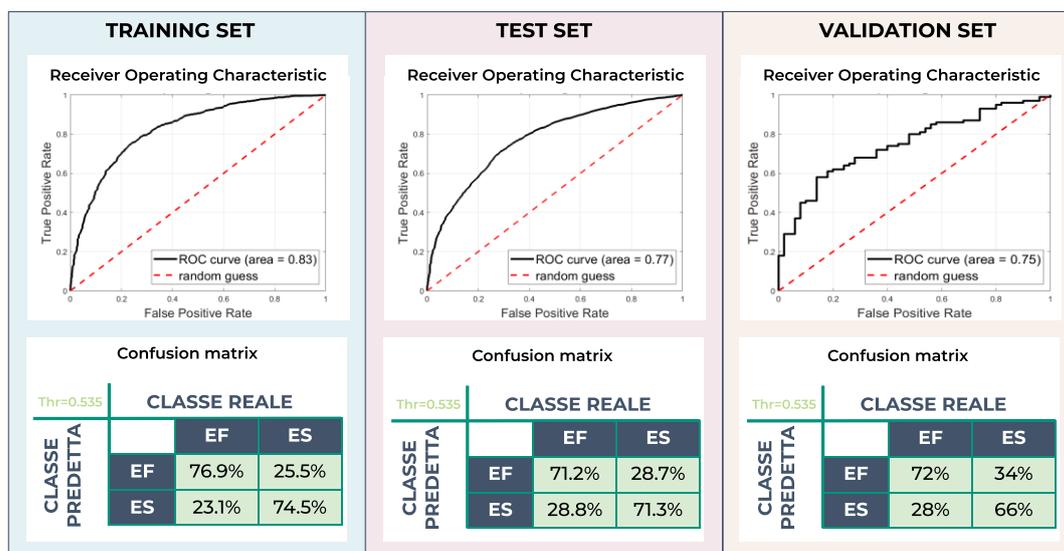


Figura 28. Confronto tra le prestazioni delle 9 topologie di rete neurale analizzate, in termini di corretta classificazione ed errori di misclassificazione. In verde è evidenziata la topologia di rete neurale individuata come migliore tra le nove analizzate.

La struttura n.6 è stata allenata solamente 10 volte a causa dell'elevato tempo computazionale necessario per completare l'apprendimento, dunque è stato rappresentato questa volta l'errore standard delle ripetizioni dell'allenamento di ogni topologia. Osservando i grafici è possibile individuare la struttura di rete che permette di avere maggiore corretta classificazione per entrambe le classi, valutando con priorità le prestazioni sul test set, maggiore capacità di generalizzare e maggiore bilanciamento tra gli errori di misclassificazione delle classi. È stata, allora, scelta l'architettura triangolare organizzata in 6 layers (struttura 6) ed è stato così allenato e testato un classificatore. In aggiunta, è stata eseguita la fase di validazione che permette di valutare se il comportamento del modello permane anche nella classificazione di nuovi pazienti. Le prestazioni del modello costruito sono state valutate in dettaglio tramite la realizzazione delle curve ROC e delle confusion matrices (figura 29).



*Figura 29. In figura sono rappresentate le curve ROC costruite tramite le predizioni del modello rispetto ai pazienti ad esso sottoposti per ciascuno dei tre sottoinsiemi (training set, test set, validation set). È possibile notare il valore della area sottesa a ciascuna curva, in particolare pari a 0.83 per quanto riguarda il training set, pari a 0.77 per quanto riguarda il test set e pari a 0.75 per quanto riguarda il validation set. Sono, inoltre, riportate le confusion matrices ottenute dalla classificazione dei pazienti appartenenti al training set, al test set e al validation set.*

## 7.2 Feature Selection: algoritmo genetico

Al fine di costruire un sistema di supporto alla decisione in ambito clinico è fondamentale controllare la quantità di informazioni inerenti un paziente richieste al personale sanitario per poter usufruire della capacità predittiva del sistema sviluppato. Infatti, un numero di variabili che caratterizzano un evento di ventilazione meccanica invasiva pari a 85 risulta difficile da sostenere dal punto di vista della pratica clinica. È quindi, importante cercare di eliminare le variabili irrilevanti o debolmente rilevanti in modo tale da ridurre il numero di misurazioni fisiologiche da effettuare al paziente. Questo comporterà anche una riduzione della complessità del sistema e dei tempi di allenamento del modello, ma è necessario selezionare un sottoinsieme di variabili che riesca a mantenere buone prestazioni del modello. Potrebbe addirittura accadere che alcune variabili non siano solamente poco informative, ma anche contengano informazioni ridondanti o siano fonte di rumore e magari eliminandole potrebbero migliorare le prestazioni del sistema. Quando si lavora con modelli di machine learning quali le reti neurali non è sempre facile eliminare le variabili irrilevanti, siccome le reti neurali possono estrarre conoscenza dai dati anche riconoscendo dei legami non lineari tra le variabili. Considerando, nel contesto delle reti neurali, altri metodi che prevedono l'eliminazione di variabili altamente correlate tra loro, è importante fare attenzione che la variabile scartata non abbia insieme ad un'altra variabile ancora un legame informativo e quindi si sta eliminando l'informazione legata al valore congiunto delle due variabili risultando, perciò, in una diminuzione delle prestazioni del classificatore. Alla luce di queste considerazioni, nel presente progetto di tesi è stata intrapresa una feature selection, associata a un'ottimizzazione dell'architettura della rete neurale, tramite un algoritmo genetico.

Gli algoritmi genetici sono metodi di ottimizzazione meta-euristici che si ispirano alla ricombinazione di due catene di DNA, ai meccanismi di selezione naturale e ai principi della genetica. Il problema di ottimizzazione consiste nel cercare di minimizzare o massimizzare una funzione di costo, chiamata fitness in riferimento alla capacità di un individuo alla sopravvivenza, cercando una soluzione ottimale tramite l'esplorazione dello spazio di ricerca. In quanto metodi meta-euristici non garantiscono di trovare l'ottimo globale, anche se potrebbero trovarlo, ma sono utili per esplorare piuttosto velocemente un numero sufficiente di soluzioni quando lo spazio di ricerca è ampio. L'algoritmo è di tipo iterativo ed è basato su una ricerca locale che si sviluppa in quattro fasi:

- **Inizializzazione.** Viene ricercato un insieme di soluzioni, dette individui, ammissibili che costituiranno la popolazione iniziale.
- **Costruzione del vicinato.** Le nuove soluzioni da valutare sono generate a partire da quelle attuali usando operazioni elementari. Essendo l'algoritmo genetico un metodo che lavora contemporaneamente con una popolazione di soluzioni che fa evolvere nel tempo, viene generato a ogni iterazione un nuovo gruppo di soluzioni.
- **Verifica dell'accettabilità.** Tramite un criterio di accettabilità si determina quali tra le soluzioni create rispettano i vincoli imposti e costituiranno la nuova popolazione di soluzioni.
- **Raggiungimento della condizione di stop.** Normalmente consiste in un numero massimo di iterazioni o nel raggiungimento della condizione di stabilità. Finché essa non è avverata l'algoritmo prosegue in una nuova iterazione, costruendo una nuova popolazione.

Per plasmare l'algoritmo genetico sul problema clinico trattato in questo progetto di tesi è stata, innanzitutto, codificata la soluzione con un insieme di 88 bit. I primi 85 bit servono a indicare per ciascuna feature la sua presenza (1) o assenza (0) nel sottoinsieme di features selezionate, mentre i restanti 3 bit indicano quale architettura di rete è stata scelta. Sono state valutate 5 possibili strutture, corrispondenti a quelle triangolari che hanno da 1 a 5 layers nascosti. I vincoli imposti alle soluzioni consistono nella presenza di un numero di variabili selezionate dalla soluzione compreso tra 10 e 30, come ragionevole compromesso tra la complessità del problema iniziale e la sostenibilità della raccolta dei dati nella pratica clinica. Inoltre, banalmente, la stringa binaria generata riferita all'architettura della rete neurale deve codificare solo per i primi 5 numeri (0-4).

Il numero di possibili soluzioni del problema, senza considerare i vincoli, corrisponde a  $2^n - 1$  avendo  $n$  bit per codificare la soluzione. Tenendo conto dei vincoli, invece, è possibile calcolare il numero totale di soluzioni ammissibili tramite il coefficiente binomiale  $n! / [(n-k)! k!]$  per tutti i  $k \in [10, 30]$  nello specifico problema trattato. Per stimare il numero di soluzioni esplorate dall'algoritmo tra tutte le possibili soluzioni del problema si tiene conto della seguente relazione, ipotizzando che vengano generate soluzioni già esplorate in iterazioni precedenti: **# individui nella popolazione + (# individui generati  $\times$  # iterazioni)**. Generalmente è preferibile avere tanti individui nella popolazione e poche iterazioni, così da avere più punti di partenza che possono evolvere verso ottimi

diversi. Però quando la funzione di fitness è pesante dal punto di vista computazionale è preferibile ridurre il numero di individui, in quanto un maggior numero di individui corrisponde a un maggior tempo computazionale dell'iterazione.

La funzione di fitness, da minimizzare, è stata in questo studio definita come: **1 - corretta classificazione totale + 0,3 × |sensibilità - specificità|**. La funzione scelta permette di valutare le prestazioni globali del classificatore allenato con i parametri dettati dalla soluzione considerata e tiene conto del bilanciamento tra sensibilità e specificità penalizzando un comportamento del modello sbilanciato tra le due classi.

Il numero di iterazioni massime dell'algoritmo è stato impostato pari a 5000 ed è stato aggiunto, come condizione di stabilità, il raggiungimento di 50 iterazioni consecutive senza riscontrare un miglioramento nella funzione di fitness. Inoltre, la popolazione è stata definita con un numero di individui pari a 5000, che si manterrà nel corso delle iterazioni essendo la popolazione a numerosità costante.

Alla prima iterazione, partendo dalla popolazione iniziale costruita, viene valutata la fitness di tutti gli individui allenando una rete neurale caratterizzata dai parametri imposti dall'individuo considerato e viene identificata la soluzione che restituisce la fitness minore. Questa operazione verrà fatta ad ogni iterazione, per i nuovi individui creati, in modo tale da trovare eventualmente una soluzione a cui corrisponde un valore di fitness minore all'ottimo temporaneamente trovato.

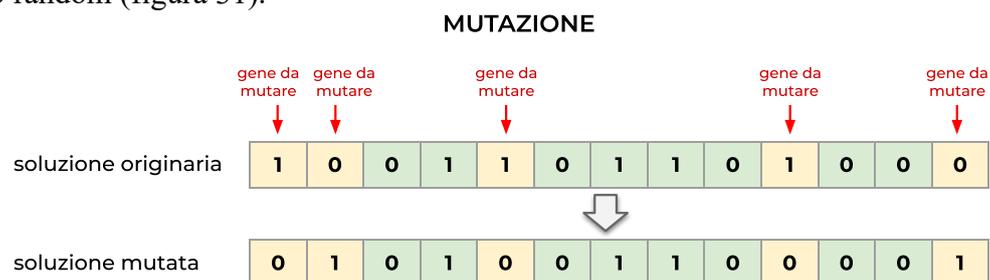
La creazione delle nuove soluzioni parte dalla selezione di un certo numero di individui dalla popolazione attuale che diventano quindi i genitori, essendo quelli che consentono di generare nuove soluzioni. In questo studio è stato scelto di estrarre un numero di genitori pari al 80% degli individui presenti nella popolazione. La generazione delle nuove soluzioni avviene attraverso l'uso di operatori genetici che si ispirano ai fenomeni di mutazione e crossover del materiale genetico. A seguito dell'eliminazione delle soluzioni non ammissibili tra quelle create e la ricerca di un nuovo ottimo temporaneo, viene effettuata la ricostruzione della popolazione a partire da quella precedente e dalle nuove soluzioni trovate, così da avere una popolazione da evolvere nell'iterazione successiva con la stessa numerosità di quella di partenza. Anche questa fase è dominata dagli operatori di selezione, che seguono lo stesso principio anche nella fase di selezione dei genitori. Gli individui vengono ordinati secondo la loro fitness e come genitori o come individui della popolazione ricostruita si selezionano le soluzioni secondo un criterio di selezione direttamente proporzionale al valore di fitness. La ricostruzione della popolazione così definita permette di inserire nella popolazione da evolvere all'iterazione successiva anche gli individui migliori della popolazione precedente.

Una volta scelti i genitori viene applicato l'operatore genetico di crossover che a partire da 2 individui genera 2 soluzioni ricombinando la coppia di genitori (figura 30).



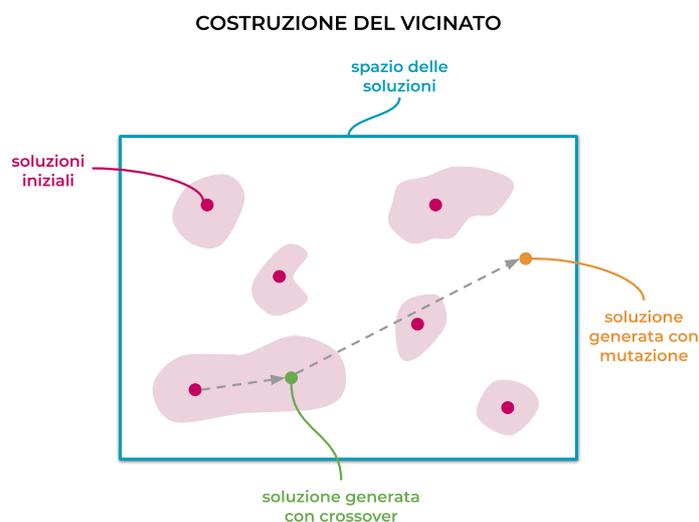
*Figura 30. Rappresentazione dell'operazione eseguita durante la fase di crossover delle soluzioni.*

Nell'algoritmo è stato impostato un numero di punti di taglio pari a 4 e la loro posizione è variabile nelle varie iterazioni e ad ogni accoppiamento. Questa operazione viene eseguita nell'algoritmo implementato con una probabilità di sottoporre le soluzioni a crossover pari a 0,9. Ai nuovi individui così creati viene, in seguito, applicato l'operatore genetico di mutazione che agisce su una singola soluzione e ne muta alcuni bit scelti in modo random (figura 31).



*Figura 31. Rappresentazione dell'operazione eseguita durante la fase di mutazione delle soluzioni.*

La mutazione è effettuata con una probabilità di mutare i geni di un individuo, cioè i bit di una soluzione, pari a 0,4 alla prima iterazione e poi tale probabilità decresce a passi di 0,01 ogni 50 iterazioni, fino a raggiungere un minimo di 0,1. Questo operatore permette di esplorare zone più lontane dello spazio di ricerca, cercando di evitare di permanere in ottimi locali (figura 32). Inoltre, la scelta di diminuire progressivamente la probabilità di mutazione nel corso delle iterazioni è motivata dal ridotto numero di individui nella popolazione rispetto alla totalità delle soluzioni possibili del problema e dunque si vogliono esplorare zone dello spazio delle soluzioni che non sono rappresentate dalla popolazione iniziale, ma poi si vuole convergere a un ottimo.



**Figura 32.** Raffigurazione della diversa esplorazione dello spazio delle soluzioni applicando l'operatore di crossover (verde) piuttosto che l'operatore di mutazione (arancione).

L'algoritmo genetico descritto è stato ripetuto per 50 volte facendo partire l'evoluzione sempre dalla stessa popolazione iniziale. L'algoritmo genetico è un metodo stocastico in quanto si basa sull'estrazione di numeri casuali e questo fa sì che partendo dalla stessa condizione iniziale la soluzione finale possa non essere la stessa. Quindi è utile sperimentare se l'algoritmo di ottimizzazione ricade sempre nello stesso ottimo, oppure in diversi ottimi locali e dunque valutare la stabilità dell'algoritmo.

In questo studio tra le 50 ripetizioni effettuate non sono state trovate soluzioni simili (cioè che differiscono per, al massimo, 2 geni) o uguali, mentre la numerosità del sottoinsieme maggiore di soluzioni che presentano una fitness uguale è pari a 4. Le 50 soluzioni trovate corrispondono, inoltre, a un valore di fitness compreso tra 0,2717 e 0,2823 mentre il range all'iterazione iniziale della funzione di fitness spaziava da 0,2827 a 0,5081. Tutte le 50 soluzioni trovate selezionano un numero di variabili compreso tra 27 e 30. Questi dettagli portano a pensare che l'algoritmo ricada in tanti ottimi locali simili tra loro. A questo punto è difficile individuare un'unica soluzione migliore tra le altre, perché non si avrebbe una scelta dettata da motivazioni forti e dunque sono state selezionate 5 soluzioni tra tutte quelle trovate. Ne è stata selezionata una per ogni architettura di rete indagabile dall'algoritmo, tra quelle che presentavano un valore di fitness più basso.

**Tabella 1.** Dettagli relativi alle 5 soluzioni individuate.

	fitness	topologia	# features selezionate
<b>Soluzione 1</b>	0.2721	struttura 1	30
<b>Soluzione 2</b>	0.2717	Struttura 2	30
<b>Soluzione 3</b>	0.2751	Struttura 3	30
<b>Soluzione 4</b>	0.2763	Struttura 4	27
<b>Soluzione 5</b>	0.2749	Struttura 5	29

È possibile osservare quali tra le 85 variabili sono state selezionate da ciascuna delle cinque soluzioni individuate, evidenziate in verde nella tabella 2. Da tale

rappresentazione è possibile notare che non è presente un pattern di variabili selezionate comune tra le cinque soluzioni. Dunque, sono presenti molteplici combinazioni di variabili che sottoposte a una rete neurale permettono di ottenere le stesse prestazioni.

Tabella 2. Variabili selezionate da ciascuna delle 5 soluzioni individuate.

Per ciascuna soluzione, al fine di valutare il comportamento medio del classificatore che definiscono, sono stati effettuati 100 addestramenti del modello e sono state, dunque, calcolate le prestazioni medie (figura 34).

CONFUSION MATRICES DELLE SOLUZIONI SELEZIONATE

	SOLUZIONE 1	SOLUZIONE 2	SOLUZIONE 3	SOLUZIONE 4	SOLUZIONE 5																																																																											
TRAINING SET	<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>69.9% ±2.8%</td><td>29.6% ±3.8%</td><td></td></tr> <tr><td>ES</td><td>30.5% ±2.8%</td><td>70.4% ±3.8%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		69.9% ±2.8%	29.6% ±3.8%		ES	30.5% ±2.8%	70.4% ±3.8%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>71.3% ±5.5%</td><td>30.5% ±3.3%</td><td></td></tr> <tr><td>ES</td><td>28.7% ±5.5%</td><td>69.5% ±3.3%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		71.3% ±5.5%	30.5% ±3.3%		ES	28.7% ±5.5%	69.5% ±3.3%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>70.3% ±3.5%</td><td>30.1% ±6.6%</td><td></td></tr> <tr><td>ES</td><td>29.7% ±3.5%</td><td>69.9% ±6.6%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		70.3% ±3.5%	30.1% ±6.6%		ES	29.7% ±3.5%	69.9% ±6.6%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>69.8% ±3.9%</td><td>30.2% ±4.7%</td><td></td></tr> <tr><td>ES</td><td>30.2% ±3.9%</td><td>69.8% ±4.7%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		69.8% ±3.9%	30.2% ±4.7%		ES	30.2% ±3.9%	69.8% ±4.7%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>71.5% ±2.9%</td><td>29.6% ±3.7%</td><td></td></tr> <tr><td>ES</td><td>28.5% ±2.9%</td><td>70.4% ±3.7%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		71.5% ±2.9%	29.6% ±3.7%		ES	28.5% ±2.9%	70.4% ±3.7%	
	100 rip		CLASSE REALE																																																																													
CLASSE PREDETTA	EF	ES																																																																														
	69.9% ±2.8%	29.6% ±3.8%																																																																														
ES	30.5% ±2.8%	70.4% ±3.8%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	71.3% ±5.5%	30.5% ±3.3%																																																																														
ES	28.7% ±5.5%	69.5% ±3.3%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	70.3% ±3.5%	30.1% ±6.6%																																																																														
ES	29.7% ±3.5%	69.9% ±6.6%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	69.8% ±3.9%	30.2% ±4.7%																																																																														
ES	30.2% ±3.9%	69.8% ±4.7%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	71.5% ±2.9%	29.6% ±3.7%																																																																														
ES	28.5% ±2.9%	70.4% ±3.7%																																																																														
TEST SET	<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>68.8% ±3.0%</td><td>27.7% ±3.4%</td><td></td></tr> <tr><td>ES</td><td>31.2% ±3.0%</td><td>72.3% ±3.4%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		68.8% ±3.0%	27.7% ±3.4%		ES	31.2% ±3.0%	72.3% ±3.4%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>69.3% ±5.4%</td><td>28.6% ±3.1%</td><td></td></tr> <tr><td>ES</td><td>30.7% ±5.4%</td><td>71.4% ±3.1%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		69.3% ±5.4%	28.6% ±3.1%		ES	30.7% ±5.4%	71.4% ±3.1%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>68.4% ±3.3%</td><td>29.3% ±6.1%</td><td></td></tr> <tr><td>ES</td><td>31.6% ±3.3%</td><td>70.7% ±6.1%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		68.4% ±3.3%	29.3% ±6.1%		ES	31.6% ±3.3%	70.7% ±6.1%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>67.9% ±3.5%</td><td>28.3% ±4.0%</td><td></td></tr> <tr><td>ES</td><td>32.1% ±3.5%</td><td>71.7% ±4.0%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		67.9% ±3.5%	28.3% ±4.0%		ES	32.1% ±3.5%	71.7% ±4.0%		<table border="1"> <tr><td colspan="2">100 rip</td><td colspan="2">CLASSE REALE</td></tr> <tr><td rowspan="2">CLASSE PREDETTA</td><td>EF</td><td>ES</td><td></td></tr> <tr><td>69.7% ±2.9%</td><td>28.2% ±3.2%</td><td></td></tr> <tr><td>ES</td><td>30.3% ±2.9%</td><td>71.8% ±3.2%</td><td></td></tr> </table>	100 rip		CLASSE REALE		CLASSE PREDETTA	EF	ES		69.7% ±2.9%	28.2% ±3.2%		ES	30.3% ±2.9%	71.8% ±3.2%	
	100 rip		CLASSE REALE																																																																													
CLASSE PREDETTA	EF	ES																																																																														
	68.8% ±3.0%	27.7% ±3.4%																																																																														
ES	31.2% ±3.0%	72.3% ±3.4%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	69.3% ±5.4%	28.6% ±3.1%																																																																														
ES	30.7% ±5.4%	71.4% ±3.1%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	68.4% ±3.3%	29.3% ±6.1%																																																																														
ES	31.6% ±3.3%	70.7% ±6.1%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	67.9% ±3.5%	28.3% ±4.0%																																																																														
ES	32.1% ±3.5%	71.7% ±4.0%																																																																														
100 rip		CLASSE REALE																																																																														
CLASSE PREDETTA	EF	ES																																																																														
	69.7% ±2.9%	28.2% ±3.2%																																																																														
ES	30.3% ±2.9%	71.8% ±3.2%																																																																														

Figura 33. Sono riportate in figura le confusion matrices, espresse in termini di valori medi e deviazioni standard, ottenute dalla classificazione dei pazienti appartenenti al training set, al test set e al validation set per ciascuna delle 5 soluzioni.

È importante porre in luce la presenza di alcune features, tra le 20 individuate da questo processo di feature selection, che sono state selezionate da tutte o quasi tutte le 50 ripetizioni effettuate dall’algoritmo genetico. In particolare, la variabile che riporta la pressione media delle vie aeree è stata scelta da tutte le soluzioni, la variabile che rappresenta la durata in giorni della somministrazione di farmaci sedativi al paziente è stata scelta da 47 soluzioni su 50 e la variabile che indica la cloremia è stata scelta 46 volte su 50. È interessante notare che queste tre variabili mostravano una percentuale di elementi mancanti molto ridotta, rispettivamente 0,9%, 0% e 0,2%. In aggiunta anche le restanti features selezionate presentavano un lieve missing degree, inferiore al 3%, tranne che per due variabili, cioè la concentrazione di albumina e di aspartato aminotransferasi nel sangue, che avevano una percentuale di elementi mancanti pari a circa la metà dei pazienti. Dunque è rilevante verificare che le features scelte attraverso il processo di features selection non siano quelle maggiormente inputate, infatti, tale osservazione permette di capire che il processo di sostituzione dei valori mancanti non ha modificato il contenuto informativo del dataset.

## Capitolo 8

### Discussione e conclusioni

L'indagine sui modelli di machine learning adottabili per affrontare il problema clinico dello svezzamento dei pazienti dal ventilatore meccanico, svolta nel presente progetto di tesi, ha portato alla costruzione di alcuni classificatori tramite l'impiego di due tecniche di apprendimento supervisionato differenti: il Light Gradient Boosting Machine e le reti neurali. Per ciascuna tecnica è stata anche affrontata la ricerca di un sottoinsieme di variabili, con numerosità inferiore alle 85 selezionate originariamente, che permettesse di raggiungere prestazioni paragonabili a quelle del modello allenato sul dataset completo della totalità delle variabili.

Le prestazioni predittive dei modelli sviluppati sono state confrontate calcolando diversi indici quantitativi di valutazione della classificazione binaria effettuata dal modello, quali l'area sottesa dalla curva ROC, l'accuratezza, la specificità, la sensibilità e il valore predittivo positivo. Per il calcolo di queste misure si ricorre al conto di alcuni elementi del set di dati considerato:

- TP. Numero di veri positivi, cioè tutti quegli eventi di ventilazione classificati dal modello come estubazioni riuscite e realmente appartenenti alla classe ES.
- TN. Numero di veri negativi, cioè tutti quegli eventi di ventilazione classificati dal modello come estubazioni fallite e realmente appartenenti alla classe EF.
- FP. Numero di falsi positivi, cioè gli eventi di ventilazione riconosciuti dal modello come estubazioni riuscite, ma che in realtà sono andate incontro a fallimento dello svezzamento.
- FN. Numero di falsi negativi, cioè gli eventi di ventilazione riconosciuti dal modello come tentativi di estubazione falliti, ma in realtà il paziente ha tollerato lo svezzamento.

Il primo indicatore calcolato è l'accuratezza, che permette di valutare la corretta classificazione totale del modello, rappresentando la capacità del modello di predire accuratamente l'esito dell'estubazione di un paziente. Indica, dunque la capacità predittiva globale rispetto a tutti gli elementi classificati, senza tenere conto dell'appartenenza degli elementi alle classi.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

I successivi indicatori valutati denotano il comportamento dei modelli nei confronti delle due classi di pazienti costituenti il dataset.

Il valore predittivo positivo permette di valutare la precisione del classificatore e dipende, infatti, dal numero di falsi positivi e dunque è indicativo della sovrastima del numero di pazienti appartenenti alla classe delle estubazioni riuscite.

$$Precision = \frac{TP}{TP+FP}$$

La recall permette, invece, di valutare la sensibilità del classificatore e dipende, infatti, dal numero di falsi negativi e dunque è indicativa della sottostima del numero di pazienti appartenenti alla classe delle estubazioni riuscite.

$$\text{Recall o Sensitivity} = \frac{TP}{TP+FN}$$

La specificità indica la probabilità che un paziente non in grado di tollerare lo svezzamento sia riconosciuto come tale e dunque un classificatore molto specifico dimostra una scarsa probabilità di misclassificare le estubazioni fallite.

$$\text{Specificity} = \frac{TN}{TN+FP}$$

È stata, inoltre, misurata l'area sottesa alla curva ROC (Receiver Operating Characteristic) calcolando l'integrale del tasso di veri positivi ( $R_{TP}$ ) rispetto al tasso di falsi positivi ( $R_{FP}$ ), essendo il primo una funzione del secondo lungo la curva.

$$AUC = \int_0^1 R_{TP}(R_{FP}) \delta R_{FP}$$

Il valore dell'area è compreso tra 0 e 1 e permette di valutare la capacità del classificatore di discernere tra la classe delle estubazioni fallite e quella delle estubazioni riuscite. Questo parametro è spesso calcolato come indicatore delle capacità predittive dei modelli costruiti negli studi presenti in letteratura che affrontano il problema clinico esaminato in questo progetto di tesi.

Nella tabella 3 sono riportati gli indicatori delle prestazioni dei modelli costruiti, che permettono di trarre alcune considerazioni.

*Tabella 3. Confronto capacità predittive dei modelli costruiti.*

		AUC	Accuracy	Precision	Sensitivity	Specificity
<b>LightGBM</b> 85 features	Training set	1.00	1.00	1.00	1.00	1.00
	Test set	0.87	0.78	0.96	0.78	0.78
	Validation set	0.89	0.77	0.92	0.72	0.88
<b>LightGBM</b> 42 features	Training set	0.99	0.99	0.99	0.99	0.99
	Test set	0.86	0.77	0.96	0.77	0.80
	Validation set	0.87	0.77	0.93	0.71	0.90
<b>ANN</b> 85 features	Training set	0.83	0.76	0.76	0.75	0.77
	Test set	0.78	0.71	0.94	0.71	0.71
	Validation set	0.75	0.68	0.83	0.66	0.72
<b>ANN</b> soluzione 1 30 features	Training set	0.75	0.70	0.68	0.77	0.63
	Test set	0.76	0.76	0.93	0.79	0.63
	Validation set	0.74	0.69	0.82	0.68	0.70
<b>ANN</b> soluzione 2 30 features	Training set	0.79	0.71	0.70	0.75	0.68
	Test set	0.78	0.75	0.93	0.77	0.66
	Validation set	0.77	0.74	0.88	0.71	0.80

<b>ANN</b> soluzione 3 30 features	Training set	0.77	0.71	0.70	0.74	0.68
	Test set	0.75	0.72	0.92	0.73	0.64
	Validation set	0.76	0.71	0.84	0.69	0.74
<b>ANN</b> soluzione 4 27 features	Training set	0.80	0.72	0.70	0.77	0.67
	Test set	0.78	0.76	0.93	0.78	0.65
	Validation set	0.77	0.73	0.86	0.71	0.76
<b>ANN</b> soluzione 5 29 features	Training set	0.80	0.73	0.72	0.76	0.71
	Test set	0.78	0.75	0.93	0.76	0.66
	Validation set	0.76	0.71	0.86	0.68	0.78

\* *LightGBM = Light Gradient Boosting Machine; ANN = Artificial Neural Network.*

Osservando i dati presenti nella tabella 3 si nota che la tecnica del Light Gradient Boosting Machine è affetta dal fenomeno dell'overfitting e dunque il modello si adatta perfettamente ai dati di training durante la fase di apprendimento, sviluppando una complessità elevata. Questo comporta una riduzione della capacità, del modello costruito, di generalizzare le informazioni apprese durante l'allenamento e dunque di mantenere le stesse prestazioni anche su un insieme nuovo di pazienti. Considerato che l'insieme di pazienti selezionati per eseguire l'allenamento è stato attentamente definito in modo tale da essere il più possibile rappresentativo della varietà di pazienti esistenti nel dataset e le strategie esistenti rivolte alla riduzione del fenomeno di overfitting sono state opportunamente adottate, la scarsa capacità di generalizzazione è riconducibile a un training set non sufficientemente numeroso. Sebbene il training set impiegato nella fase di addestramento comprende un totale di 2300 pazienti, la documentazione studiata suggerisce un numero di esempi su cui addestrare il modello ancora maggiore. Nonostante questo fenomeno le prestazioni mostrate dal modello in fase di testing sono comparabili con quelle mostrate in fase di validazione, quindi in questo caso mostra una buona capacità di generalizzazione. Il lightGBM raggiunge, comunque, capacità predittive generalmente superiori a quelle delle reti neurali.

Per quanto riguarda il comportamento delle reti neurali ottenute tramite la strategia di ricerca dell'algoritmo genetico, si osserva che esse hanno delle prestazioni comparabili, sebbene siano fondate sull'analisi di differenti variabili per riuscire a discernere il quadro clinico di pazienti che non tollerano l'estubazione da quello di pazienti il cui svezzamento va a buon fine. Potrebbe, perciò, essere possibile costruire un classificatore che sia il risultato della combinazione delle cinque reti neurali individuate. In questo modo la predizione combinata dei pazienti misclassificati da una singola rete neurale potrebbe rivelarsi corretta e quindi, le prestazioni di tale modello potrebbero essere superiori a quelle delle singole reti neurali. Inoltre, osservando che le soluzioni trovate tramite l'algoritmo genetico tendono a saturare il numero massimo di variabili consentito nella strategia di ricerca, si potrebbe provare ad aumentare tale soglia, consentendo all'algoritmo di individuare un ottimo, magari migliore a quelli attualmente trovati, caratterizzato dalla presenza di un numero di variabili leggermente superiore a 30.

È possibile, inoltre, notare che in generale i modelli sviluppati sono molto precisi e dunque quando stabiliscono che un paziente è in grado di tollerare l'estubazione molto probabilmente sarà così, siccome i falsi positivi sono rari.

Grazie all'analisi degli studi esistenti in letteratura che affrontano il problema clinico dello svezzamento dei pazienti dalla ventilazione meccanica invasiva è possibile effettuare un confronto dei risultati ottenuti nel presente progetto di tesi con quelli ottenuti dagli studi che sfruttano gli stessi modelli di machine learning e la stessa tipologia di pazienti [13, 14, 15]. Si può, così, riscontrare che le prestazioni dei modelli approfonditi nel presente studio sono comparabili a quelle dei modelli sviluppati in altri studi.

*Tabella 4. Confronto tra le capacità predittive dei modelli costruiti nel presente progetto di tesi e in altri studi esistenti in letteratura.*

	<b>AUC</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Sensitivity</b>	<b>Specificity</b>
LightGBM 85 features	0.89	0.77	0.92	0.72	0.88
LightGBM 42 features	0.87	0.77	0.93	0.71	0.90
LightGBM Chen et al. 2019 68 features	0.81	0.80	-	0.75	0.83
LightGBM Chen et al. 2019 36 features	0.82	0.80	-	0.75	0.84
ANN 85 features	0.75	0.68	0.83	0.66	0.72
ANN soluzione 1 30 features	0.74	0.69	0.82	0.68	0.70
ANN soluzione 2 30 features	0.77	0.74	0.88	0.71	0.80
ANN soluzione 3 30 features	0.76	0.71	0.84	0.69	0.74
ANN soluzione 4 27 features	0.77	0.73	0.86	0.71	0.76
ANN soluzione 5 29 features	0.76	0.71	0.86	0.68	0.78
ANN Chen et al. 2019 68 features	0.68	0.79	-	0.82	0.63
ANN Chen et al. 2019 36 features	0.74	0.80	-	0.80	0.69
ANN Hsieh et al. 2018 37 features	0.85	-	0.96	0.81	-
ANN Kuo et al. 2015 17 features	0.83	0.80	-	0.82	0.73

\* *LightGBM = Light Gradient Boosting Machine; ANN = Artificial Neural Network.*

Esaminando i valori riportati nella tabella 4 si può evidenziare, per quasi tutti i classificatori, una maggiore specificità rispetto alla sensibilità e dunque i modelli tendono a riconoscere meglio i pazienti che andranno incontro a un fallimento dell'estubazione.

La ricerca svolta nel presente progetto di tesi ha permesso di riconoscere e analizzare i modelli e le variabili adatte ad affrontare il problema clinico dello svezzamento di un paziente dalla ventilazione meccanica invasiva. Tale studio ha l'intenzione di essere proseguito con la costruzione di una Datawarehouse orientata a un'indagine prospettica, che raccolga i parametri di pazienti intubati in condizioni critiche che saranno ammessi, durante il periodo dedicato alla raccolta, al reparto di terapia intensiva dell'azienda ospedaliero-universitaria 'Città della Salute e della Scienza di Torino'. Parallelamente allo studio condotto nel progetto di tesi è stata portata avanti una richiesta al comitato etico per ottenere l'accesso ai dati presenti in cartelle cliniche di pazienti ammessi al reparto di terapia intensiva in anni passati. I risultati ottenuti nel presente progetto di tesi sono un riferimento per la scelta dei parametri da estrarre dalle cartelle cliniche dei pazienti e analogamente per la definizione dei parametri da misurare nella raccolta prospettica, così come anche per la determinazione dei pazienti da arruolare nello studio prospettico e da selezionare tra quelli presenti nell'archivio delle ammissioni in terapia intensiva. Inoltre, sono stati analizzati gli aspetti relativi all'applicazione dei due modelli di machine learning nell'ambito della classificazione dei pazienti intubati in condizioni critiche a seconda dell'esito dell'estubazione. Queste analisi sono utili a fornire una percezione del comportamento dei modelli nell'estrazione di conoscenza dai dati di pazienti intubati in condizioni critiche e dunque eseguire una costruzione più consapevole del classificatore che verrà effettivamente impiegato come strumento di supporto nella decisione, da parte del medico intensivista, di intraprendere un tentativo di estubazione in un paziente. Sebbene i dati appartenenti ai pazienti ricoverati all'ospedale di Boston possono essere aggregati a quelli dei pazienti dell'ospedale torinese, sarebbe ottimale considerare anche le caratteristiche di altre popolazioni differenti per la modellizzazione del problema reale. Questo permetterebbe di ottenere una maggiore generalizzabilità dello studio effettuato e del sistema CAD sviluppato.

Ulteriori riflessioni sono emerse durante lo svolgimento del lavoro di tesi, che potrebbero essere in prosieguo sviscerate. La prima riflessione riguarda la componente temporale del quadro clinico di un paziente che potrebbe essere considerata nella costruzione del classificatore, in modo tale da tenere conto dell'evoluzione di certi parametri nella distinzione di pazienti che andranno incontro a fallimento o successo dello svezzamento dal ventilatore meccanico. A questo proposito, non saranno più considerate solamente le ultime registrazioni dei parametri effettuate prima del tentativo di estubazione del paziente o il valore minimo, massimo e medio dei parametri vitali durante il periodo di intubazione, ma anche il loro decorso. Per quanto riguarda la definizione di estubazione fallita alcuni studi considerano un periodo più lungo a seguito dell'estubazione nel quale rilevare una eventuale re-intubazione. Ad esempio lo studio condotto da Hsieh et al nel 2018 [13] ritiene estubazioni fallite quei casi in cui il paziente è stato re-intubato o è deceduto nelle 72 ore successive all'estubazione. Tale gruppo di ricerca ricorda, inoltre, che l'utilizzo di una ventilazione non invasiva (NIV) a seguito dell'estubazione potrebbe porre la necessità di re-intubare il paziente. Relativamente ai criteri di esclusione applicati per definire la popolazione di pazienti da esaminare nello studio è stato pensato che i pazienti che hanno sperimentato un'auto-estubazione potrebbero essere, invece, inseriti nella popolazione reclutata. Questa considerazione è nata dal fatto che una buona

percentuale di pazienti che si liberano volontariamente dal tubo endotracheale durante il periodo di svezzamento non necessita di una re-intubazione. Dunque, potrebbe essere utile valutare anche tali pazienti nella coorte di pazienti reclutati in modo tale da analizzare anche il quadro clinico di questi pazienti. Questo fenomeno ricorda ancora che alcuni pazienti sono sottoposti al trattamento di ventilazione meccanica invasiva più a lungo del necessario. Un'ultima riflessione fa emergere l'idea che potrebbe essere utile suddividere i pazienti in categorie, ad esempio tramite l'impiego di reti SOM (Self Organizing Map), e costruire un classificatore per ciascuna categoria, ottenendo così dei classificatori specializzati per una particolare tipologia di pazienti, che magari saranno più performanti.

In conclusione, i risultati ottenuti nel presente progetto di tesi avvalorano l'idea che un modello di machine learning sia adatto a predire la probabilità dei pazienti intubati in condizioni critiche di tollerare la respirazione autonoma e a discernere, perciò, quelli che andranno incontro al fallimento del tentativo di estubazione da quelli che sperimenteranno lo svezzamento con successo. I risultati ottenuti incoraggiano, quindi, la prosecuzione dello studio al fine di costruire un sistema CAD in grado di supportare il medico intensivista nell'individuazione del momento ottimale per svezzare il paziente dal ventilatore meccanico.

## Appendice A

### Studio delle variabili impiegate in letteratura

1: variabile utilizzata  
 0: variabile non registrata e non utilizzata  
 -1: variabile presente nel dataset, ma non utilizzata per allenare il modello

	articolo 1	articolo 2	articolo 3	articolo 4	articolo 5	articolo 6	articolo 7	articolo 8
<b>DEMOGRAPHIC DATA</b>								
Age (years)	1	1	0	0	0	1	1	1
Sex (Male/Female)	1	-1	0	0	0	1	1	1
BMI (Kg/m <sup>2</sup> )	1	-1	0	0	0	0	1	0
Weight	0	0	0	1	1	1	0	0
Height	0	0	0	0	0	1	0	0
Weight loss	0	0	0	0	0	1	0	0
>= 65 years old	0	0	0	0	0	0	0	1
<b>Hospital stay (days)</b>								
Hospital stay (days)	0	-1	0	0	0	0	1	0
ICU stay (days)	0	-1	0	0	0	0	1	0
Internal medicine	0	0	0	0	0	0	0	1
<b>CLINICAL SCORES</b>								
APACHE II	1	1	0	0	0	0	1	1
TISS scale	1	0	0	0	0	0	1	1
Glasgow Coma Scale	1	0	0	0	0	1	1	1
SOFA Sequential Organ Failure Assessment	0	0	0	0	0	1	0	0
<b>COMORBIDITIES</b>								
Chronic kidney disease	0	0	0	0	0	1	0	1
Uremia	0	0	0	0	0	0	1	0
Chronic lung disease	1	0	0	0	0	0	0	0
Chronic liver disease	1	0	0	0	0	1	0	1
Liver cirrhosis	0	0	0	0	0	0	1	0
Chronic hemo dialysis	1	0	0	0	0	0	0	0
Cardiovascular accident	1	0	0	0	0	0	0	0
Old stroke	1	0	0	0	0	0	1	0
Chronic cardiac insufficiency	0	0	0	0	0	1	0	0
Coronary heart disease	0	0	0	0	0	0	1	1
Arrhythmia	0	0	0	0	0	1	0	0
Valvular disease	0	0	0	0	0	1	0	0
Pulmonary heart disease	0	0	0	0	0	1	0	0
Peripheral vascular	0	0	0	0	0	1	0	0
Hypertension	0	0	0	0	0	1	0	0
Neurological disease	0	0	0	0	0	1	0	1
Chronic Obstructive Pulmonary Disease	0	0	0	0	0	1	1	0
Diabetes uncomplicated	0	0	0	0	0	1	0	0
Diabetes complicated	0	0	0	0	0	1	0	0
Diabetes	1	0	0	0	0	0	1	1
Hypothyroidism	0	0	0	0	0	1	0	0
Peptic ulcer	0	0	0	0	0	1	0	0
AIDS	0	0	0	0	0	1	0	0
Lymphoma	0	0	0	0	0	1	0	0
Metastatic tumors	0	0	0	0	0	1	0	0
Solid tumor/ active cancer	1	0	0	0	0	1	1	1
Rheumatoid arthritis	0	0	0	0	0	1	0	0
Coagulopathy	0	0	0	0	0	1	0	0
Obesity	0	0	0	0	0	1	0	0
Blood loss Anemia	0	0	0	0	0	1	0	0
Deficiency anemia	0	0	0	0	0	1	0	0
Alcohol abuse	0	0	0	0	0	1	0	0
Drug abuse psychoses	0	0	0	0	0	1	0	0

Depression	0	0	0	0	0	1	0	0
Electrolyte and fluid disorder	0	0	0	0	0	1	0	0
Psychoses	0	0	0	0	0	1	0	0
Immunocompromised	0	0	0	0	0	0	0	1
Respiratory disease	0	-1	0	0	0	0	0	1
Cardiovascular disease	0	-1	0	0	0	0	0	0
Neuromuscular disease	0	-1	0	0	0	0	0	0
<b>CLINICAL INTERVENTION</b>								
Sedation days	0	0	0	0	0	1	0	0
Vasopressor	0	0	0	0	0	1	0	0
<b>ETIOLOGY OF INTUBATION / REASON FOR INITIATING MV</b>								
Acute COPD exacerbation	0	1	0	0	0	0	0	1
Pneumonia	0	1	0	0	0	0	0	1
Heart failure	0	1	0	0	0	0	0	0
Neuromuscular disease	0	1	0	0	0	0	0	0
Hypoventilation	0	0	0	0	0	0	0	1
Airway obstruction	0	0	0	0	0	0	0	1
Cardiogenic pulmonary edema	0	0	0	0	0	0	0	1
Septic shock	0	0	0	0	0	0	0	1
Post-operation	0	0	0	0	0	0	0	1
<b>ETIOLOGY OF RESPIRATORY FAILURE</b>								
Pulmonary system	0	0	0	0	0	0	0	1
Cardiovascular system	0	0	0	0	0	0	0	1
Neurological system	0	0	0	0	0	0	0	1
Renal system	0	0	0	0	0	0	0	1
Gastrointestinal system	0	0	0	0	0	0	0	1
<b>WEANING PARAMETERS</b>								
RSBI Rapid Shallow Breathing Index	1	-1	0	0	1	-1	0	1
Maximum Inspiratory Pressure (cm H <sub>2</sub> O)	1	-1	1	1	1	1	0	1
Maximum Expiratory Pressure (cm H <sub>2</sub> O)	1	0	0	0	0	0	0	1
Mechanical ventilation duration (hours)	1	1	0	0	0	1	0	1
Tidal volume V <sub>T</sub> (mL of ideal body weight)	0	-1	1	1	1	1	0	0
Breathing frequency (breaths/min)	0	-1	0	0	0	0	0	0
Minute ventilation	0	0	0	0	0	1	0	0
Mean Airway Pressure	0	0	0	0	0	1	0	0
<b>WEANING METHODS</b>								
T-piece	0	0	0	0	0	0	0	1
Pressure support<= 8cmH <sub>2</sub> O	0	0	0	0	0	0	0	1
SBT	0	0	0	0	0	0	0	1
<b>NUMBER OF ATTEMPTS WEANING</b>								
T-piece	0	0	0	0	0	0	0	1
Pressure support<= 8cmH <sub>2</sub> O	0	0	0	0	0	0	0	1
SBT	0	0	0	0	0	0	0	1
<b>PRE-EXTUBATION DATA</b>								
FiO <sub>2</sub>	1	-1	0	1	1	0	0	1
Pressure support level (cm H <sub>2</sub> O)	1	0	0	0	0	0	0	1
PEEP Positive End-Expiratory Pressure (cm H <sub>2</sub> O)	1	0	0	0	0	0	0	1
Minute ventilation (or respiratory minute volume) (L/min)	1	0	1	1	1	0	0	1
Work Of Breathing	0	0	0	0	1	0	0	0
min Heart Rate (beats/min)	0	0	0	0	0	-1	0	0
max Heart Rate (beats/min)	0	0	0	0	0	-1	0	0
mean Heart Rate (beats/min)	1	-1	0	1	1	1	0	1
min Respiratory Rate (breaths/min)	0	0	0	0	0	-1	0	0
max Respiratory Rate (breaths/min)	0	0	0	0	0	1	0	0
mean Respiratory Rate (breaths/min)	1	1	1	0	1	-1	0	1
min Temperature	0	0	0	0	0	-1	0	0
max Temperature	0	0	0	0	0	-1	0	0
mean Temperature	0	0	0	0	0	1	0	0

min Systolic blood pressure (mmHg)	0	0	0	0	0	-1	0	0
max Systolic blood pressure (mmHg)	0	0	0	0	0	-1	0	0
mean Systolic blood pressure (mmHg)	0	-1	0	0	0	1	0	0
min Diastolic blood pressure (mmHg)	0	0	0	0	0	-1	0	0
max Diastolic blood pressure (mmHg)	0	0	0	0	0	-1	0	0
mean Diastolic blood pressure (mmHg)	0	-1	0	0	0	1	0	0
min Arterial pressure	0	0	0	0	0	-1	0	0
max Arterial pressure	0	0	0	0	0	-1	0	0
mean Arterial pressure	1	0	0	1	0	1	0	1
min Percutaneous oxygen saturation	0	0	0	0	0	1	0	0
max Percutaneous oxygen saturation	0	0	0	0	0	-1	0	0
mean Percutaneous oxygen saturation	0	0	0	0	0	-1	0	0
SpO <sub>2</sub> (%) before SBT	0	-1	0	0	0	0	0	0
SpO <sub>2</sub> (%) after SBT	0	-1	0	0	0	0	0	0
mean Inspiratory time during SBT	0	1	0	1	0	0	0	0
mean Expiratory time during SBT	0	1	0	0	0	0	0	0
change in RSBI after 30 min during SBT	0	-1	0	0	0	0	0	0
RSBI after 30 min after SBT	0	-1	0	0	0	0	0	0
Spontaneous breathing/ Ventilatory rate	0	1	0	1	0	0	0	0
spontaneous Tidal volume V <sub>T</sub> (mL)	0	1	0	0	1	0	0	0
<b>BLOOD ANALYSIS and BGA</b>								
arterial pH	1	0	0	1	0	1	0	1
PaCO <sub>2</sub> (mmHg)	1	0	0	1	0	1	0	1
PaO <sub>2</sub> (mmHg)	1	0	0	1	1	1	0	1
PaO <sub>2</sub> /FIO <sub>2</sub>	1	0	0	0	1	0	0	1
SaO <sub>2</sub>	0	0	0	1	1	0	0	0
lag time between last blood gas result and extubation	0	0	0	1	0	0	0	0
aA ratio (alveolar-arterial gradient)	0	0	0	0	1	0	0	0
CVO <sub>2</sub> central vein oxygenated	0	0	0	0	0	-1	0	0
Hemoglobin (g/dL)	1	0	0	0	0	1	0	1
Hematocrit (%)	1	0	0	0	1	0	0	1
Blood urea nitrogen (mg/dL)	1	0	0	0	0	1	0	1
Blood glucose	0	0	0	0	0	1	0	0
Creatinine (mg/dL)	1	0	0	0	0	1	0	1
Base excess	0	0	0	0	0	1	0	0
Sodium (meq/L)	1	0	0	0	0	1	0	1
Potassium (meq/L)	1	0	0	0	0	1	0	1
Calcium (mg/dL)	1	0	0	0	0	1	0	1
Lactic acid	0	0	0	0	0	1	0	0
Chloride	0	0	0	0	0	0	0	0
Phosphate (mg/dL)	1	0	0	0	0	0	0	1
Albumin (g/dL)	1	0	0	0	0	-1	0	1
Monocytes	0	0	0	0	1	0	0	0
Neutrophils	0	0	0	0	1	0	0	0
Lymphocytes	0	0	0	0	1	0	0	0
White blood cell	0	0	0	0	0	1	0	0
Platelet	0	0	0	0	0	1	0	0
Troponin	0	0	0	0	0	-1	0	0
total Protein	0	0	0	0	0	-1	0	0
Brain Natriuretic Peptide	0	0	0	0	0	-1	0	0
C-reactive protein	0	0	0	0	0	-1	0	0
GPT Glutamic-pyruvic transaminase	0	0	0	0	0	-1	0	0
GOT Glutamic oxalacetic transaminase	0	0	0	0	0	-1	0	0
total Bilirubin	0	0	0	0	1	-1	0	0
Prothrombin time	0	0	0	0	0	1	0	0
Activated partial thromboplastin time APTT	0	0	0	0	0	1	0	0
International Normalized Ratio	0	0	0	0	0	1	0	0

Gli articoli a cui si fa riferimento nella precedente tabella sono i seguenti.

**articolo 1**

An Artificial Neural Network model for predicting successful extubation in Intensive Care Units. 2018. Hsieh et al. [13]

**articolo 2**

Improvement in the prediction of ventilator weaning outcomes by an Artificial Neural Network in a medical ICU. 2015. Kuo et al. [14]

**articolo 3**

A comparison of human and machine predictions of successful weaning from mechanical ventilation. 2000. Gottschalk et al. [16]

**articolo 4**

Machine learning to predict extubation outcome in premature infants. 2013. Mueller et al. [17]

**articolo 5**

Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database. 2012. Mikhno et al. [11]

**articolo 6**

Prediction of extubation failure for Intensive Care Unit patients using High Gradient Boosting Machine. 2019. Chen et al. [15]

**articolo 7**

Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in Intensive Care Units. 2018. Hsieh et al. [18]

**articolo 8**

Predicting weaning difficulty for planned extubation patients with an Artificial Neural Network. 2019. Hsieh et al. [19]

## Appendice B

### Variabili estratte o calcolate dal database MIMIC-III

	Categoria	Tipologia	Momento della misurazione	Missing degree (%)
Age (years)	dati demografici	discreta	giorno dell'ammissione in ICU	0
Gender	dati demografici	binaria	-	0
Weight (kg)	dati demografici	continua	giorno dell'intubazione	17,4
Weight loss (kg)	dati demografici	continua	differenza tra il giorno dell'intubazione e dell'estubazione	67,8
Height (cm)	dati demografici	continua	-	32,3
Glasgow Coma Scale	punteggi di gravità	discreta	primo giorno di ricovero in ICU	2,8
SOFA score	punteggi di gravità	discreta	primo giorno di ricovero in ICU	0
Sedation days (days)	somministrazione di vasopressori e sedativi	continua	dal primo giorno di ricovero in ICU al momento dell'estubazione	0
Vasopressor	somministrazione di vasopressori e sedativi	binaria	periodo di ventilazione meccanica invasiva	0
Chronic cardiac insufficiency	comorbidità	binaria	ammissione in ospedale	0
Arrhythmia	comorbidità	binaria	ammissione in ospedale	0
Valvular disease	comorbidità	binaria	ammissione in ospedale	0
Pulmonary heart disease	comorbidità	binaria	ammissione in ospedale	0
Peripheral vascular	comorbidità	binaria	ammissione in ospedale	0

Hypertension	comorbidità	binaria	ammissione in ospedale	0
Stroke	comorbidità	binaria	ammissione in ospedale	0
Other neurological	comorbidità	binaria	ammissione in ospedale	0
Chronic obstructive pulmonary disease	comorbidità	binaria	ammissione in ospedale	0
Diabetes uncomplicated	comorbidità	binaria	ammissione in ospedale	0
Diabetes complicated	comorbidità	binaria	ammissione in ospedale	0
Hypothyroidism	comorbidità	binaria	ammissione in ospedale	0
Chronic kidney disease	comorbidità	binaria	ammissione in ospedale	0
Liver disease	comorbidità	binaria	ammissione in ospedale	0
Peptic ulcer	comorbidità	binaria	ammissione in ospedale	0
AIDS	comorbidità	binaria	ammissione in ospedale	0
Lymphoma	comorbidità	binaria	ammissione in ospedale	0
Metastatic tumors	comorbidità	binaria	ammissione in ospedale	0
Solid tumors	comorbidità	binaria	ammissione in ospedale	0
Rheumatoid arthritis	comorbidità	binaria	ammissione in ospedale	0
Coagulopathy	comorbidità	binaria	ammissione in ospedale	0
Obesity	comorbidità	binaria	ammissione in ospedale	0
Blood loss anemia	comorbidità	binaria	ammissione in ospedale	0
Deficiency anemia	comorbidità	binaria	ammissione in ospedale	0
Alcohol abuse	comorbidità	binaria	ammissione in ospedale	0
Drug abuse psychoses	comorbidità	binaria	ammissione in ospedale	0

Depression	comorbidità	binaria	ammissione in ospedale	0
Electrolyte and fluid disorders	comorbidità	binaria	ammissione in ospedale	0
Psychoses	comorbidità	binaria	ammissione in ospedale	0
Mechanical ventilation duration (hours)	parametri ventilatori	continua	ultimo valore registrato prima dell'estubazione	0
Tidal volume (mL/B)	parametri ventilatori	continua	ultimo valore registrato prima dell'estubazione	0,4
Minute ventilation (L/min)	parametri ventilatori	continua	ultimo valore registrato prima dell'estubazione	0,6
Mean Airway Pressure (cmH2O)	parametri ventilatori	continua	ultimo valore registrato prima dell'estubazione	0,9
Positive inspiratory pressure (cmH2O)	parametri ventilatori	continua	ultimo valore registrato prima dell'estubazione	0,8
mean Heart rate (bpm)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
max Heart rate (bpm)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
min Heart rate (bpm)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
mean Respiratory rate (insp/min)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
max Respiratory rate (insp/min)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
min Respiratory rate (insp/min)	segnali vitali	continua	periodo di ventilazione	0,01

			meccanica invasiva	
mean Temperature (°C)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,1
max Temperature (°C)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,1
min Temperature (°C)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,1
mean Systolic blood pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
max Systolic blood pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
min Systolic blood pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
mean Diastolic blood pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
max Diastolic blood pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
min Diastolic blood pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
mean Mean arterial pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
max Mean arterial pressure (mmHg)	segnali vitali	continua	periodo di ventilazione	0,02

			meccanica invasiva	
min Mean arterial pressure (mmHg)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,02
mean Percutaneous oxygen saturation SpO2 (%)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
max Percutaneous oxygen saturation SpO2 (%)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
min Percutaneous oxygen saturation SpO2 (%)	segnali vitali	continua	periodo di ventilazione meccanica invasiva	0,01
White blood cells (K/uL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,3
Hemoglobin (g/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,1
Platelet (K/uL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,2
Sodium (mEq/L)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,1
Calcium (mg/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	9
Potassium (mEq/L)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,1
Lactic acid (mmol/L)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	20,3
Chloride (mEq/L)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,2

Blood glucose (mg/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,1
Creatinine (mg/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,1
Blood Urea Nitrogen (mg/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	0,1
Troponin (ng/mL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	74,8
total Protein (g/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	96,7
Brain natriuretic peptide (pg/mL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	96,7
C-reactive protein (mg/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	98
total Bilirubin (mg/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	48,9
Albumin (g/dL)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	52,2
GPT (IU/L)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	48,8
GOT (IU/L)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	48,9
arterial pH	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	2,4
PaCO2 (mmHg)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	2,7
PaO2 (mmHg)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	2,7

Base excess (mEq/L)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	2,7
Central vein oxygenated (%)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	28,1
Prothrombin Time (s)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	3,6
activated Partial Thromboplastin Time (s)	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	3,7
International Normalized Ratio	analisi ematiche e emogasanalisi	continua	ultimo valore registrato prima dell'estubazione	3,6

## Bibliografia

- [1] Survival of mechanically ventilated patients admitted to intensive care units.  
*Taha Ismaeil, Jawaher Almutairi, Rema Alshaikh, Zahrah Althobaiti, Yassin Ismaiel and Fatmah Othman.* Saudi Med J. 2019 Aug; 40(8): 781–788.
- [2] Patients on weaning trials classified with support vector machines.  
*Ainara Garde, Rico Schroeder, Andreas Voss, Pere Caminal, Salvador Benito and Beatriz F Giraldo.* 2010 Physiol. Meas. 31 979.
- [3] Epidemiology of Weaning Outcome according to a New Definition. The WIND Study.  
*B'eduneau, Pham, Schortgen, et al.* American Journal of Respiratory and Critical Care Medicine Volume 195 Number 6 | March 15 2017
- [4] Long-term outcomes and health care utilization after prolonged mechanical ventilation.  
*Hill A.D., Fowler R.A., Burns K.E., Rose L., Pinto R.L., Scales, D.C.* Ann. Am. Thorac. Soc. 2017, 14, 355 – 362.
- [5] Mechanical ventilation weaning issues can be counted on the fingers of just one hand: part 1.  
*Luigi Vetrugno, Giovanni Maria Guadagnin, Alessandro Brussa, Daniele Orso, Eugenio Garofalo, Andrea Bruni, Federico Longhini and Tiziana Bove.* Ultrasound Journal. 2020
- [6] Ventilator discontinuation protocols.  
*Haas, C.F., Loik, P.S.* Respir. Car. 2012, 57, 1649 – 1662
- [7] Evidence-based guidelines for weaning and discontinuing ventilatory support: a collective task force facilitated by the American College of Chest Physicians; the American Association for Respiratory Care; and the American College of Critical Care Medicine.  
*MacIntyre NR, Cook DJ, Ely EW Jr, Epstein SK, Fink JB, Heffner JE, et al.* Chest 2001;120(6):375S-395S. Also in: Respir Care 2002;47(2): 69-90.
- [8] Parameter selection for and implementation of a web-based decision-support tool to predict extubation outcome in premature infants.  
*Hulsey Thomas C, Knapp Rebecca G, Annibale David J, Wagner Carol L, Mueller Martina, Almeida Jonas S.* BMC Medical Informatics and Decision Making, 01 March 2006, Vol.6(1), p.11
- [9] MIMIC-III, a freely accessible critical care database.  
*Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG.* Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Disponibile al link: <http://www.nature.com/articles/sdata201635>

- [10] A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units.  
*Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, Barbara E Engelhardt.* 2017.
- [11] Prediction of Extubation Failure for Neonates with Respiratory Distress Syndrome Using the MIMIC-II Clinical Database.  
*Arthur Mikhno, Colleen M Ennett.* 34th Annual International Conference of the IEEE EMBS San Diego, California USA. 2012.
- [12] Prognostic factors and outcomes of unplanned extubation.  
*Chien-Ming Chao, Mei-I. Sung, Kuo-Chen Cheng, Chih-Cheng Lai, Khee-Siang Chan, Ai-Chin Cheng, Shu-Chen Hsing, Chin-Ming Chen*Chien-Ming. Springer Nature. 2017.
- [13] An artificial neural network model for predicting successful extubation in intensive care units.  
*Meng-Hsuen Hsieh, Meng-Ju Hsieh, Chin-Ming Chen, Chia-Chang Hsieh, Chien-Ming Chao, Chih-Cheng Lai.* J Clin Med. 2018 Aug 25;7(9):240.
- [14] Improvement in the prediction of ventilator weaning outcomes by an artificial neural network in a medical ICU.  
*Hung-Ju Kuo, Hung-Wen Chiu, Chun-Nin Lee, Tzu-Tao Chen, Chih-Cheng Chang, Mauo-Ying Bien.* Respir Care 2015 Nov;60(11):1560-9.
- [15] Prediction of extubation failure for intensive care unit patients using high gradient boosting machine.  
*Tingting Chen, Jun Xu, Haochao Ying, Xiaojun Chen, Ruiwei Feng, Xueling Fang, Honghao Gao, Jian Wu.* IEEE 11 October 2019.
- [16] A comparison of human and machine predictions of successful weaning from mechanical ventilation.  
*Gottschalk et al.* Medical Decision Making 20(2):160-9 April 2000.
- [17] Machine learning to predict extubation outcome in premature infants.  
*Martina Mueller, Carol C. Wagner, Romesh Stanislaus, Jonas S. Almeida.* The 2013 International Joint Conference on Neural Networks (IJCNN). Dallas, TX, USA. 4-9 Aug. 2013.
- [18] Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in Intensive Care Units.  
*Hsieh, M.H., Hsieh, M.J., Chen, CM. et al.* Sci Rep **8**, 17116. 2018.
- [19] Predicting weaning difficulty for planned extubation patients with an Artificial Neural Network.  
*Hsieh MH, Hsieh MJ, Cheng AC, Chen CM, Hsieh CC, Chao CM, Lai CC, Cheng KC, Chou W.* Medicine (Baltimore). 2019 Oct;98(40):e17392.