Xi *et al.*
Health Information Science and Systems (2025) 13:48
https://doi.org/10.1007/s13755-025-00364-4

Health Information Science and Systems

RESEARCH

K-cql: an arterial blood gas analysis-based deep offline reinforcement learning algorithm for mechanical ventilation treatment

Jiaying Xi[1], Shaojie Dong[2], Haoquan Zhou[3] and Yunbo Zhao[1,2*]

Abstract

Mechanical ventilation is employed as a supportive therapy for patients with respiratory failure, but the optimal ventilator settings for patient are often unknown and rely on manual adjustment by physicians. Improper parameter settings may lead to severe complications such as lung injury. To personalize mechanical ventilation and predict the optimal ventilator parameters for patients, we propose a ventilator parameter tuning algorithm. This algorithm integrates clinical expertise in ventilator tuning via Arterial Blood Gas (ABG) analysis with data-driven methods. We perform K-means clustering algorithm on patient dataset based on ABG values for the first time, and the classified data was used to train a deep offline reinforcement learning model based on conservative Q-learning (CQL), therefore we named it the K-CQL algorithm. The introduction of human expert knowledge improves the effectiveness of the entire model. Our evaluation based on Fitted Q Evaluation (FQE) on the MIMIC-III dataset shows that the expected return of the output strategy of K-CQL is 1.76 times that of the physicians, and more importantly, the introduction of intermediate rewards related to ABG analysis further improves it. We also demonstrated that the algorithm is capable of recommending mechanical ventilation parameters within a safe range according to clinical nursing standards.

Keywords: Mechanical ventilation, Arterial blood gas analysis, Deep offline reinforcement learning, Conservative q-learning, K-means

Introduction

Respiratory failure is caused by multiple factors leading to dysfunction in pulmonary ventilation and gas exchange, which, if not treated in time, can result in multi-organ failure and even life-threatening consequences [1]. Thus, symptomatic and causal treatments are required based on clinical symptoms and test results, including the use of mechanical ventilation to control the disease. The optimal ventilator settings vary between individuals and are generally unknown [2], requiring manual adjustments by physicians based on the patient's condition. The accuracy of these adjustments is directly influenced by the physician's knowledge and experience, and incorrect parameter tuning can worsen the patient's condition or even lead to death. For instance, improper ventilator settings may cause ventilator-induced lung injury, diaphragmatic dysfunction, pneumonia, and oxygen toxicity [3]. To prevent such complications and provide optimal care, personalized mechanical ventilation is essential. Integrating cutting-edge artificial intelligence (AI) technologies to explore suitable strategies for ventilator adjustments has become a critical research focus at the intersection of medicine and engineering.

Traditionally, recommendations for ventilator parameters in medical settings have relied on automatic control principles. For example, [4] applied adaptive fuzzy sliding mode control (AFSMC) to the respiratory system to aid patients with respiratory distress; [5] used a proportional-integral (PI) controller to develop a closed-loop respiratory pacemaker model. However, these traditional approaches often exhibit limited adaptability when it comes to adjusting parameters. The application of reinforcement learning in the medical field shows a good

*Correspondence: ybzhao@ustc.edu.cn
[2] Institute of Advanced Technology, University of Science and Technology of China, Shushan District, Hefei 230031, Anhui, China
Full list of author information is available at the end of the article

prospect [6, 7, 9]. More recent research has proposed the use of machine learning to personalize mechanical ventilation treatments, with a focus on supervised learning and reinforcement learning (RL). In [10], a bayesian classification model was used to classify patient's illness, followed by an artificial neural network (ANN) to output parameters such as frequency, tidal volume, and fraction of inspired oxygen. However, deep supervised learning, while allowing feature extraction, overlooks the sequential nature of mechanical ventilation. Furthermore, supervised learning methods can only imitate physician's decision, which may lead to suboptimal treatment. RL, on the other hand, interacts with the environment and receives immediate feedback from patients in the form of rewards, considering the accumulation of future discounted rewards [11–13], thereby improving physician's strategies and is widely used in the medical field [6].

Recent studies have shown that RL holds great potential in mechanical ventilation. It has been proven to be suitable for solving decision-making problems in time-series data and can potentially address ventilator parameter tuning [14, 15]. Peine et al. [16] described the ventilator management process as a Markov decision process (MDP) and employed the Q-learning algorithm to recommend ventilator parameters. Chen et al. [15] designed a simulated environment using a long short-term memory (LSTM) network and used the soft actor-critic (SAC) algorithm to provide clinical decision support for ventilator parameter adjustments. Kondrup et al. [17] introduced DeepVent, an offline deep RL model based on Conservative Q-learning (CQL) to predict optimal ventilator parameters, and compare it with Double Deep Q Network (DDQN). In [18] , transformer was combined with CQL [19] to propose a model capable of diagnosing patient's conditions and predicting optimal ventilator parameters. Zhang et al. [1] were the first to consider safety issues in ventilator parameter recommendations by identifying the optimal policy from a fixed dataset.

However, to the best of our knowledge, previous work on modeling ventilator adjustments using MDP has primarily remained theoretical, without focusing on indicators of Arterial Blood Gas (ABG) analysis. In real clinical settings, ABG analysis indicators are highly correlated with the gas exchange process [20], physicians adjust ventilator parameters based on it, as ABG analysis provides better guidance for ventilator adjustments [21]. For instance, [22] developed a mathematical model based on ABG analysis to assess patients' conditions and analyze gas exchange during ventilation. Therefore, we embedded a K-means clustering model into the RL decision-making algorithm, incorporating prior knowledge from physicians who use ABG analysis to adjust ventilator parameters. By clustering patients based on ABG analysis before training, this approach optimizes treatment strategies and further improves therapeutic outcomes.

Our main contributions can be summarized as follows:

- Pre-classifying patients based on ABG analysis indicators is proposed for the first time, combining clinical knowledge from physicians adjusting ventilators with data-driven algorithms, thereby incorporating human prior knowledge. Specifically, a ventilator adjustment model algorithm (termed K-CQL) using K-means clustering based on ABG analysis and CQL algorithm is developed. Additionally, intermediate rewards related to ABG analysis indicators is defined to achieve faster convergence.
- The performance of the K-CQL algorithm, the DeepVent algorithm (currently the most popular deep RL model for ventilator adjustment), and the physician's strategies from the MIMIC-III dataset was evaluated and compared using Fitted Q Evaluation (FQE) based on publicly available MIMIC-III dataset [23]. The results demonstrated that the output policy of K-CQL algorithm achieved the highest expected return value, which is 1.19 times that of the DeepVent algorithm and 1.76 times that of the physicians, thereby exhibiting superior performance.

Related works

Algorithms for ventilation optimization

At present, methods for ventilation optimization in hospitals typically rely on Proportional-Integral-Derivative (PID) control, which is known to be suboptimal [24]. Siu et al. [25]designed a closed-loop adaptive controller to automatically adjust settings of ventilator based on analysis of arterial carbon dioxide. Ai et al. [5]introduced a closed-loop respiratory pacemaker, using a PI controller to adapt to various ventilation conditions. However, traditional automatic control methods evaluate a limited number of data features and are inefficient in handling high-dimensional and large-scale clinical data [26]. Machine learning (ML) methods have demonstrated their superiority in processing high-dimensional and large datasets, making them effective tools in the medical field for handling extensive clinical data. Zhu et al. [27] applied machine learning methods such as k-nearest neighbors (KNN), logistic regression, decision trees, and extreme gradient boosting to predict mortality in ventilated patients. Similarly, [10, 28] used deep neural networks to calculate outputs for frequency, tidal volume, and fraction of inspired oxygen. However, these machine learning approaches primarily focus on the relationship between ventilator parameter changes and patient

vital signs, without considering the long-term effects of ventilators on patients. Compared to classical machine learning methods, RL algorithms learn from interactions with the environment to maximize long-term rewards [29]. RL algorithms can simulate how clinicians optimize ventilator parameters through continuous interaction with patients, demonstrating potential to surpass clinical standards and providing strong evidence for RL's application in this context. Additionally, RL's mechanism of maximizing cumulative rewards enhances its performance in solving optimization problems that account for long-term effects [11]. RL have shown promising results in off-policy evaluation for predicting continuous ventilator parameters. Peine et al. [16] used Q-learning to optimize ventilator settings such as positive end expiratory pressure, fraction of inspired oxygen, and tidal volume. Kondrup et al. [17]were the first to combine deep learning with RL to adjust ventilator parameters. Chen et al. [15] designed a simulated environment using a long short-term memory (LSTM) network and utilized the soft actor-critic (SAC) algorithm to provide recommendations. Yuan et al. [18] combined transformer and CQL to propose a model capable of diagnosing conditions of patient and predicting optimal ventilator settings. Zhang et al. [1] applied RL algorithms to identify optimal strategies from a fixed dataset and were the first to address safety issues in ventilator parameter recommendations, ensuring improved efficacy and safety during treatment.

Models for parameter tuning

[16, 17] selected several physiological variables including demographics, vital signs, lab values, and fluids as the state space in their modeling of MDP, with [17] incorporating changes in Acute Physiology and Chronic Health Evaluation II (APACHE-II) scores as part of the reward function. Clinically, physicians adjust ventilator parameters based on ABG analysis, making it necessary to develop a more clinically relevant RL-based ventilator adjustment model to enhance practical applicability. Chen et al. [15] considered ABG analysis states when modeling the MDP but did not appropriately define intermediate reward values. Wang et al. [22] developed a mathematical model based on ABG analysis to determine whether the patient's ABG analysis values were within normal ranges, serving as a measure of whether goals of mechanical ventilation were achieved. Ma et al. [30] used data-driven and knowledge-driven machine learning methods to predict organ failure in intensive care unit (ICU) patients in real time, incorporating clinical prior knowledge. Based on these considerations, we embedded a classification model into the RL decision-making algorithm. By classifying patients based on ABG analysis indicators before training, the prior knowledge of physicians using ABG analysis to adjust ventilator parameters is integrated, that is to say, the professional knowledge of physicians is effectively combined with data-driven algorithms to optimize treatment strategies and further improve treatment effects.

Background

Reinforcement learning

In a RL problem, we typically model it as a MDP, represented by the tuple (S, A, P, R, γ), where S denotes the state space, A denotes the action space, P represents the transition probability matrix, R and γ are the reward function and discount factor, respectively. A policy is defined as $\pi : S \rightarrow A$ and is trained to maximize the expected cumulative discounted reward in the MDP:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(a_t|s_t))\right] \tag{1}$$

Generally, a Q-value function is defined to represent the expected cumulative reward:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(a_t|s_t))\middle| s, a\right] \tag{2}$$

Q-Learning is a classic method that trains the Q-value function by minimizing the Bellman error on Q:

$$Q \leftarrow \arg\min_{Q} \mathbb{E}\left[R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right] \tag{3}$$

Offline reinforcement learning and conservative Q-learning

Traditional RL often refers to online RL. In this scenario, the agent needs to continually interact online with the environment to obtain feedback. However, this approach is not permitted in healthcare settings, as it may put patients at risk. Offline RL algorithms can learn effective policies from previously collected static datasets without the need for further interactions, making them highly applicable in clinical settings. However, in practice, offline RL presents a major challenge: standard offline RL methods may suffer from Q-value overestimation due to distributional shift between the dataset and the learned policy. In healthcare settings, this overestimation may translate into unsafe recommendations, putting patients at risk. CQL aims to address these limitations by learning a conservative Q-function, ensuring that the expected value of the policy is lower-bounded by the true value under this Q-function. CQL achieves this by introducing a conservative term that penalizes actions deviating significantly from the data distribution, thus preventing

overestimating. Specifically, it attempts to underestimate the Q-values of out-of-distribution (OOD) state-action pairs, thereby discouraging the agent from entering OOD states. Therefore, the optimization objective of CQL becomes:

$$Q \leftarrow \arg\min_{Q} \alpha \cdot \left[\mathbb{E}_{s\sim\mathcal{D},a\sim\pi(a|s)}Q(s,a) - \mathbb{E}_{s,a\sim\mathcal{D}}Q(s,a)\right] + \frac{1}{2}\left(R(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right)^2 \quad (4)$$

The parameter α is a weighting hyperparameter used to control the strength of the penalty term. CQL suppresses the overestimation of actions that are infrequent in the dataset by maximizing the Q-values for actions $a \sim \mathcal{D}$ observed in the data, while minimizing the Q-values for actions with high Q-values under the learned policy π.

Batch-constrained Q-learning

Similar to CQL, Batch-Constrained Q-learning (BCQ) [31] was proposed to address the distributional shift problem in offline reinforcement learning. BCQ primarily focuses on constraining the policy's reliance on unseen state-action pairs from the offline dataset to mitigate the risks associated with distributional shift. When learning a policy directly from a static dataset, BCQ attempts to restrict the policy's action choices to align more closely with the observed behaviors, thus avoiding unsafe decisions that may arise from selecting actions that deviate from the data distribution. It introduces a generative model (such as a Variational Autoencoder, VAE) to generate candidate actions that are similar to those observed in the data, from which actions are selected to ensure that the policy does not significantly deviate from the historical distribution:

$$p(a \mid s) \approx \mathrm{VAE}(a|s) \quad (5)$$

$$a = \arg\max_{a' \in \mathcal{A}} Q(s,a') \cdot 1(a' \in \mathrm{VAE}(a|s)) \quad (6)$$

$1(a' \in \mathrm{VAE}(a|s))$ is a constraint that ensures only actions with a high probability in the generative model are selected.

This design effectively reduces the risk associated with out-of-distribution actions.

K-means

K-means is an unsupervised learning clustering algorithm widely applied in various data analysis and classification tasks across different fields. Its primary objective is to partition the samples in a dataset into k distinct clusters, such that the samples within the same cluster are as similar as possible, while the differences between samples in different clusters are maximized [32]. The core idea of the K-means algorithm is to minimize the total sum of squared errors (SSE) within clusters. The workflow of the algorithm is as follows:

(1) Initialization: Randomly select k points as the initial cluster centroids;
(2) Cluster assignment: Assign each sample to the nearest cluster based on the distance between the sample and the cluster centroid;
(3) Centroid update: Recompute the centroid of each cluster by averaging the positions of all samples within the cluster;
(4) Iteration: Repeat steps 2 and 3 until the centroids converge or a predefined stopping criterion is met.

The objective function of K-means is to minimize SSE within the clusters.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (7)$$

where k is the number of clusters, C_i represents the i-th cluster, μ_i is the centroid of the i-th cluster and x is a sample point in cluster C_i.

By iteratively adjusting the position of the cluster centroids and reassigning the samples, the algorithm minimizes the intra-cluster variance, achieving the optimal clustering outcome.

Method

An overview of the model architecture is here in Fig. 1. This figure illustrates a framework for ventilator parameter optimization that integrates expert knowledge with offline reinforcement learning. The overall pipeline begins with the analysis of patients' ABG indices, which are used to cluster patients into distinct subgroups with similar physiological characteristics via the K-means algorithm. This clustering process serves as a mechanism for embedding clinical prior knowledge into the modeling process. For each identified patient cluster, a separate reinforcement learning model is trained using the CQL algorithm, enabling personalized policy learning tailored to specific clinical profiles. CQL is particularly well-suited for offline settings, as it constrains the learned Q-values to avoid overestimation for out-of-distribution actions, thereby enhancing the safety and stability of the learned policy. Each model is trained on retrospective physician demonstration data, which comprising patient states, actions, and rewards. By incorporating domain knowledge through clustering and leveraging the robustness of CQL, the framework aims to generate individualized ventilator adjustment strategies that align more closely with

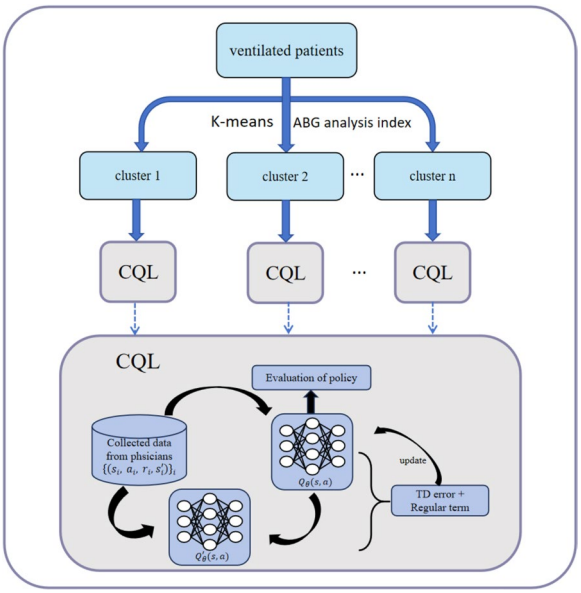

Fig. 1 Model architecture

clinical reasoning, ultimately facilitating safer and more effective mechanical ventilation treatment.

Data extraction and pre-processing

We utilized the MIMIC-III database, an open-access database containing data from the Beth Israel Deaconess Medical Center between 2001 and 2012. Standardized Query Language (SQL) was employed to extract patient data into tables with four-hour time windows. For each patient, the following data were extracted: demographics, vital signs, lab values, fluids, and mechanical ventilation settings. The first 72 h of mechanical ventilation data were selected, and patient data were divided into states, actions, and rewards. For data imputation, similar to the approach taken by Kondrup, a hybrid method was used. If less than 30% of the data was missing, k-nearest neighbors (KNN) with k=3 was applied for imputation. For missing data between 30% and 95%, a time window sampling and hold method was employed, using the initial value to replace subsequent values until a new value was reached or the limit was met. When the initial value was missing, mean imputation was performed. Finally, for variables with more than 95% missing data, they were removed from the state space. After processing, the patient dataset consisted of 19,780 samples, including 36 state variables.

Model of classification

K-means clustering is a distance-based algorithm that assigns samples to different clusters by minimizing the within-cluster sum of squared errors [33]. The algorithm iteratively adjusts the clustering centre to optimize the clustering results and selects an appropriate number of clusters to capture the underlying structure of the data. By using K-means clustering, we can categorize patients based on results of ABG analysis, with each cluster representing groups of patients sharing similar characteristics in terms of blood gas parameters. The features of each group provide insights into the physiological states of different patients, serving as a foundation for developing personalized treatment strategies. Therefore, K-means clustering plays a pivotal role in our methodology, acting as a bridge between traditional medical knowledge and advanced machine learning algorithms, thereby advancing the development of intelligent medical technologies.

We extracted relevant personal characteristics and effective blood gas analysis parameters as clustering features for the patients: age, gender, weight, spo2, ph, paco2, base excess, bicarbonate.

RL definition

The definition of our MDP is similar to the work of Peine, with the episode spanning from the patient's intubation to the subsequent 72 h.

The state space consists of 36 variables:

- Demographics: Age, gender, weight, readmission to the ICU, Elixhauser score;
- Vital Signs: SOFA, SIRS, GCS, heart rate, sysBP, diaBP, meanBP, shock index, temperature, spo2;
- Lab Values: Potassium, sodium, chloride, glucose, bun, creatinine, magnesium, carbon dioxide, Hb, WBC count, platelet count, ptt, pt, inr, pH, partial pressure of carbon dioxide, base excess, bicarbonate;
- Fluids: Urine output, vasopressors, intravenous fluids, cumulative fluid balance.

The action space comprises three ventilator settings:

- Volume of air in and out with each breath adjusted by ideal weight (Vt);
- Positive End Expiratory Pressure (PEEP);
- Fraction of Inspired Oxygen (FiO2).

The action space A is the Cartesian product of these three settings, therefore, an action is represented as a tuple $a = (v, o, p)$ with $v \in V_t$, $o \in F_iO_2$, and $p \in PEEP$.

The primary objective of our agent is to stabilize the patient's ABG analysis indicators within the normal range and ensure long-term survival. In setting the reward function, we build upon the work of Kondrup et al., which defines a terminal reward $r(s_t, a_t, s_{t+1})$ where the value is -1 if the patient dies within 90 days, and +1 otherwise. Since it is well-known that relying solely on

sparse terminal rewards can lead to suboptimal RL task performance and inefficiencies in sample utilization [34]. APACHE-II is a widely used severity-of-disease scoring system designed to assess the condition of patients in the ICU [35]. The scoring system combines various physiological parameters, age, and the patient's chronic health status to predict the mortality risk of critically ill patients. Given the importance of blood gas analysis for ventilator adjustments, we emphasize the inclusion of blood gas indicators in shaping the intermediate reward, building upon the APACHE-II score. The APACHE-II scores based on relevant physiological indicators are calculated according to a predetermined range and then summed up. For instance, regarding body temperature:

- A temperature within the range of 36.0—38.4°C is assigned 0 points.
- A temperature ranging from 38.5 to 38.9°C or from 34.0 to 35.9°C is assigned 1 point.
- A temperature within the range of 32.0—33.9°C is assigned 2 points.
- A temperature ranging from 39.0 to 40.9°C or from 30.0 to 31.9°C is assigned 3 points.
- A temperature $\geq 41.0\,^{\circ}\mathrm{C}$ or $\leq 29.9\,^{\circ}\mathrm{C}$ is assigned 4 points.

Given that ABG values constitute critically important indicators for adjusting ventilator settings, these ABG values are extracted for separate scoring and assigned a weighting factor of 2 in the APACHE-II calculation.

$$r = \begin{cases} +1 & \text{if } t+1 = l_i \text{ and } m^i_{t+1} = 1 \\ -1 & \text{if } t+1 = l_i \text{ and } m^i_{t+1} = 0 \\ \frac{2(b^i_{t+1} - b^i_t) + (c^i_{t+1} - c^i_t)}{\max_A - \min_A} & \text{otherwise} \end{cases} \tag{8}$$

where b^i_t is the Apache II score related to blood gas analysis of patient i at timestep t; c^i_t is the Apache II score except blood gas analysis of patient i at timestep t, $m^i_t = 0$ if patient i is dead at timestep t and 1 otherwise, l_i is the length of patient i's stay at the ICU, and max_A, min_A are respectively the maximum and minimum possible values of our modified Apache II score.

Experiment

Baseline

We employed five baseline methods: the physician policy, the DeepVent model, the DeepVent+ model, the K-BCQ model ,the K-DDQN model and the K-CQL-model. The physician policy comprised all transitions (s_t, a_t, s_{t+1}) observed in the dataset, thereby represented the choices made by physicians treating patients in the MIMIC-III dataset. The DeepVent model, which served as the deep RL baseline, was the most advanced deep RL model currently used for ventilator adjustment. The DeepVent+ model extended the original DeepVent by incorporating ABG-based reward shaping, enabling the evaluation of whether reward shaping can enhance the performance of the model. The K-CQL- model builds upon the DeepVent model by incorporating clustering but does not include blood gas analysis-related intermediate rewards. It was used as the baseline for the reward function to evaluate whether adding blood gas analysis-related intermediate rewards can improve the performance of the model. The comparison of the modules between these three models and the K-CQL model is presented in Table 1. The K-BCQ model and the K-DDQN model can verify the effectiveness of CQL. Compared to BCQ, CQL incorporates a conservative mechanism that underestimates the Q-values for unobserved actions, thus preventing the policy from over-optimizing unreliable actions. The conservative Q-value estimation effectively mitigates Q-value overestimation, making the policy more safe.

Hyperparameters

In order to verify the advancement of K-means clustering in this method, we added a comparative experiment and compared it with some classical and representative clustering methods, including Gaussian Mixture Model (GMM), Agglomerative Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Spectral Clustering, the results were shown in Table 2. A Silhouette Score closer to 1, a higher Calinski-Harabasz Score, and a lower Davies-Bouldin Score indicate tighter intra-cluster cohesion and greater inter-cluster separation, signifying superior clustering performance.

The Silhouette Coefficient is one of the metrics used to evaluate the quality of clustering, particularly in K-means clustering. It helps determine the effectiveness of the clustering results [36]. A high Silhouette Coefficient indicates that K-means successfully clusters the data under the current number of clusters, with data points tightly grouped within clusters and well-separated between clusters. The Silhouette Coefficient $s(i)$ for each data point i is defined as follows:

Table 1 Comparison of the modules of DeepVent, DeepVent+, K-CQL- and K-CQL. ✗: module not included; ✓: module included

Model	DEEPVENT	DEEPVENT+	K-CQL-	K-CQL
K-means	✗	✗	✓	✓
ABG-based intermediate rewards shaping	✗	✓	✗	✓

Table 2 Values of Silhouette Score, Calinski–Harabasz Score, and Davies–Bouldin Score for the five clustering methods

Clustering model	K-means	GMM	Agglomerative	DBSCAN	Spectral
Silhouette Score	0.468	0.389	0.373	− 0.345	0.237
Calinski-Harabasz Score	16637.743	16582.070	15699.460	2090.266	7848.249
Davies-Bouldin Score	0.889	0.895	0.937	2.048	0.848

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

$a(i)$ is the average distance between data point i and all other points within the same cluster; $b(i)$ is the average distance between data point i and the nearest other cluster.

By calculating the average silhouette coefficient for different cluster numbers ranging from 2 to 10, as shown in Fig. 2, we found that the silhouette coefficient was highest when the number of clusters was 3, indicating the best clustering performance at this point. Therefore, we chose 3 as the optimal number of clusters. The patient sample size was 19,780, divided into three clusters with sizes of 5,728, 8,194, and 5,858, respectively. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technology for data dimensionality reduction and visualization, primarily used to project high-dimensional data into lower-dimensional spaces (typically two or three dimensions) for a more intuitive understanding of data structure. Its core goals is to preserve local proximity relationships in the high-dimensional space, ensuring that similar points remain close in the low-dimensional space. The clustering results are shown in Fig. 3.

Regarding the selection of parameters in deep reinforcement learning, the patient dataset was divided into 80% for the training set and 20% for the validation set. Based on the parameter grid search strategy outlined in the work of Kondrup et al., the optimal values for CQL are determined as $\gamma = 0.75$, $\eta = 1 \times 10^{-6}$ and $\alpha = 0.1$. The optimal architecture consisted of two hidden layers, each with 256 units and ReLU activation functions. The models were run on an NVIDIA GeForce RTX 3090 GPU. Eeah model was then run for 2 million steps across

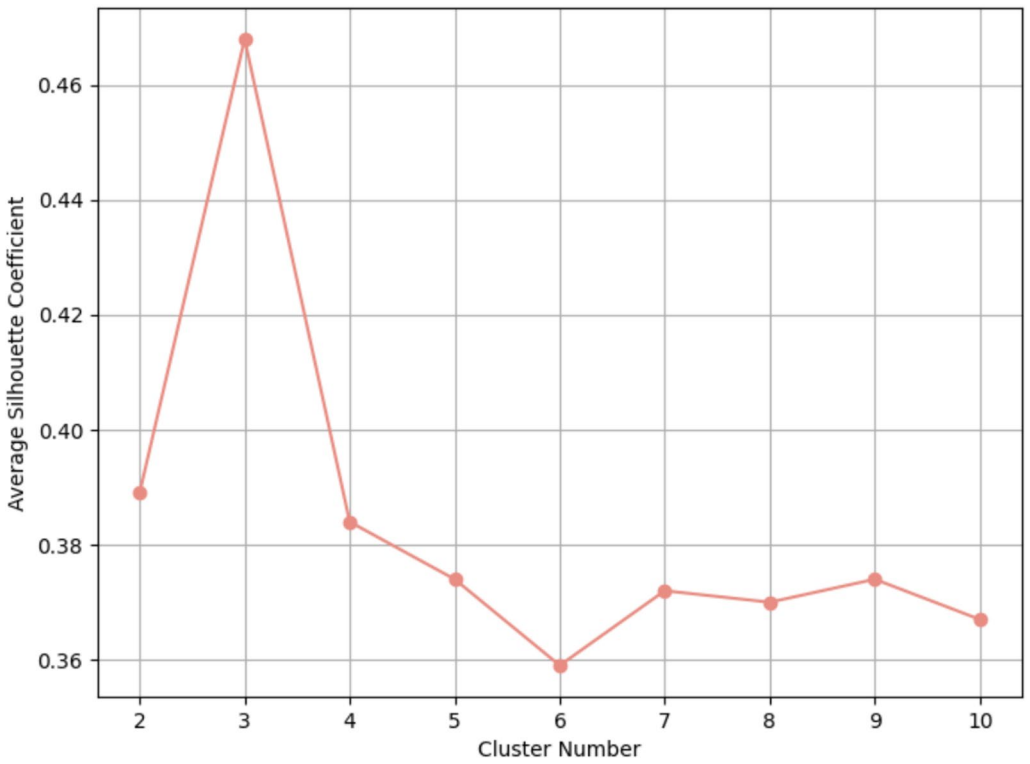

Fig. 2 Plot of silhouette coefficient as a function of the number of clusters. A higher silhouette coefficient indicates that the K-means algorithm achieves better clustering performance for the current number of clusters

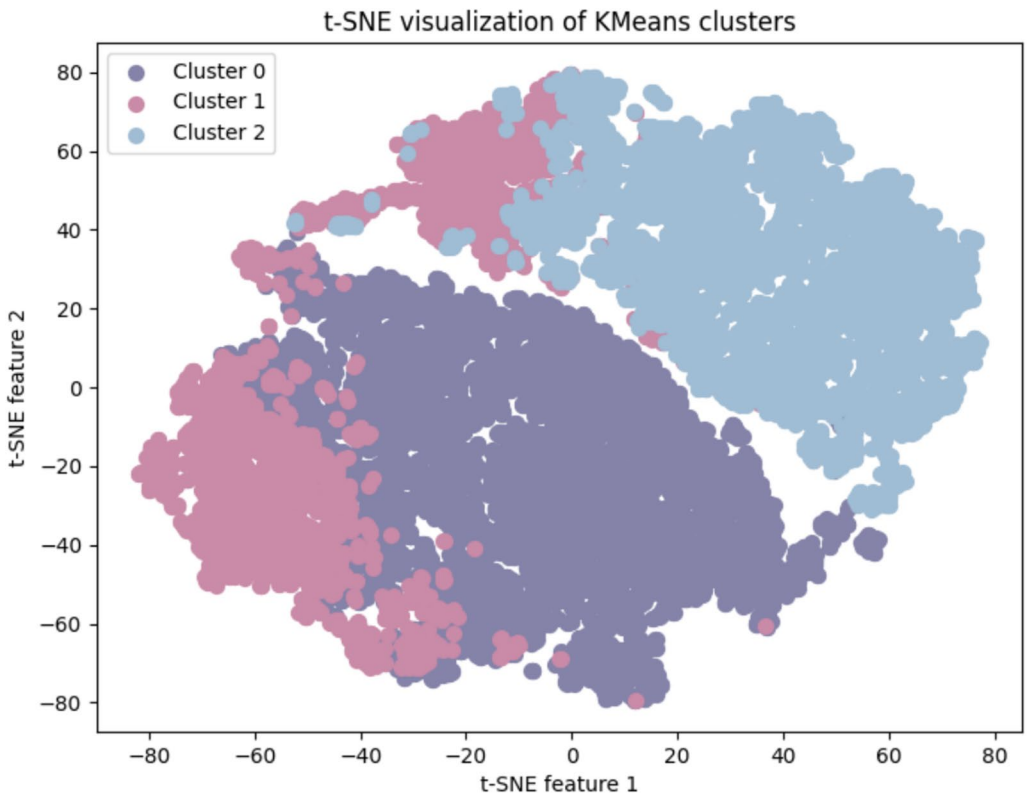

Fig. 3 Plot of clustering results. Downgrade multi-dimensional data to two-dimensional data for visual clustering

five iterations with seed numbers 42, 43, 44, 45, and 46 respectively, and the results were averaged. Each training session took approximately 16 h to 25 h to complete, with slight variations in runtime across different algorithms.

Evaluation of performance

In online RL, policies are typically evaluated through interactions with the environment. However, when the environment involves real patients in a medical context, evaluating policies in this online manner poses significant risks. Therefore, this study employed off-policy evaluation (OPE) to assess the policy using the dataset. The performance of these methods has recently been evaluated in medical settings, with FQE consistently provided the most accurate results [37]. Based on the current policy and offline data, FQE utilizes a fitted model to iteratively update the Q-values, allowing it to better approximate the true Q-function. Once the Q-function has converged, the performance of the policy can be evaluated by computing the expected value of the Q-values. Specifically, for each state, the Q-value corresponding to the optimal action is selected, and the average of these Q-values is calculated to obtain the overall performance of the policy. The performance of a policy can then be computed by taking the mean initial state value, where the initial state represents the first four hours of ventilation. Since the physician's policy effectively generates the episodes in the dataset, the cumulative discounted reward for each initial state can be calculated based on the episodes starting from that state.

Result

We first conducted ablation experiments on K-CQL to demonstrate the effectiveness of both K-means and ABG-based intermediate rewards shaping, then we investigated the performance of the K-CQL algorithm using FQE and compared it with other advanced models, demonstrating that K-CQL achieved superior performance. Subsequently, we analyzed the action distribution suggested by K-CQL, which indicated that the selected actions aligned with clinical recommendations. Finally, we further evaluated our model in out-of-distribution (OOD) scenarios, showing that K-CQL maintained high performance when applied to OOD patients, making it a safer option for real world.

Overall performance

We first conducted ablation experiments to demonstrate the effectiveness of K-means and ABG-based intermediate rewards shaping. The four algorithms presented

in Table 2 serve as comparative models in the ablation experiments. Then, we compared the performance of K-CQL-, K-CQL, K-DDQN, K-BCQ, physicians, DeepVent+ and DeepVent based on FQE (Table 3), with a 95% confidence level. The performance of K-CQL was obtained by averaging the results of deep offline reinforcement learning training performed separately on the three clusters of data.

The performance of K-CQL- is currently 1.10 times that of the state-of-the-art DeepVent. When intermediate rewards related to Arterial Blood Gas analysis are incorporated, the factor increases to 1.19. Compared to DeepVent, both the incorporation of K-means and the introduction of ABG-based intermediate rewards shaping further enhance the model's performance. The evaluation value of the K-DDQN algorithm is too small, while the K-BCQ algorithm yields negative evaluation values, both algorithms cannot adapt to this scenario. Therefore, our results indicate that the performance of K-CQL is significantly superior to that of both physician and DeepVent.

Distribution of suggested actions

Next, we evaluated the action distribution of K-CQL in comparison to DeepVent, K-BCQ, K-DDQN and human physicians. PEEP is frequently set to 5 cmH2O, but it can be personalized based on changes in physiological parameters [38]. As shown in Fig. 4, K-CQL made most of its recommendations in the 0–5 cmH2O range. Higher PEEP settings are significantly associated with an increased risk of barotrauma and pneumothorax [39], this situation should be prevented. K-CQL aligned with clinical care standards regarding FiO2, making recommendations similar to those of physicians in the MIMIC-III dataset, particularly in the 35-50% and >55% ranges. Lastly, for the optimal weight-adjusted tidal volume, the recommended range is typically within the 4-8 ml/kg range [40]. K-CQL made most of its options within the 5-10 ml/kg range. The action distribution of K-DDQN is relatively uniform, making it clearly unsuitable as an output policy. Although the output policies of both K-BCQ and DeepVent align with clinical standards, they did not achieve as high mean initial state values as K-CQL, as shown in Table 2. To sum up, we observed that K-CQL was able to provide safe recommendations based on clinical care standards for patients when compared to physician strategies.

Performance in OOD

BCQ and CQL are two commonly used algorithms for offline reinforcement learning, both designed to learn policies from offline data while mitigating the issue of Q-value overestimation.

Q-value overestimation is a frequent problem in offline reinforcement learning, which, in clinical settings, could lead to unsafe parameter recommendations for patients. Therefore, we investigated whether the recommendations made by K-BCQ and K-CQL could combat the overestimation of OOD state-action pairs. Similar to the work of Kondrup, we generated an OOD dataset to explore the overestimation issue in reinforcement learning. Outlier patients were defined as those whose state features at the initiation of mechanical ventilation fell within the top and bottom 1% of the distribution. Approximately 25% of the patients were classified as outliers. We here compute the mean initial Q values for K-BCQ and K-CQL estimated by FQE trained on our dataset, both in and out of distribution.

Since the maximum return without intermediate rewards in our dataset is set to 1, and FQE was trained on this dataset, any value above 1 should be considered an overestimation. As shown in Fig. 5, K-BCQ overestimated the values in both ID and OOD settings, and this overestimation was exacerbated in the OOD setting, indicating that the BCQ algorithm does not completely eliminate overestimation. CQL addresses Q-value overestimation more directly and systematically than BCQ. The core idea of CQL is to actively suppress overestimation, thus avoiding these issues. The average initial state value estimates of K-CQL remained below the overestimation threshold of 1 in both settings, with little variation in the OOD setting, demonstrating the stability of model across both scenarios.

Conclusion

In this paper, we propose a knowledge and data-driven approach to achieve adaptive ventilator parameter adjustment for the treatment of patients with respiratory failure, aiming to realize personalized treatment. We introduce K-CQL algorithm model, which is based on ABG analysis clustering and offline deep reinforcement learning, for individualized ventilator parameter adjustment. Our validation results, based on FQE, demonstrate that K-CQL outperforms both physicians and the DeepVent model. More importantly, incorporating

Table 3 Mean initial state value estimates for physician, DeepVent, DeepVent+, K-DDQN, K-BCQ, K-CQL- and K-CQL

Model	PHYSICIAN	DEEPVENT	DEEPVENT+	K-DDQN	K-BCQ	K-CQL-	K-CQL
Value	0.502 ± 0.007	0.743 ± 0.005	0.765 ± 0.004	0.050 ± 0.014	− 157.030 ± 2.551	0.814 ± 0.001	0.883 ± 0.003

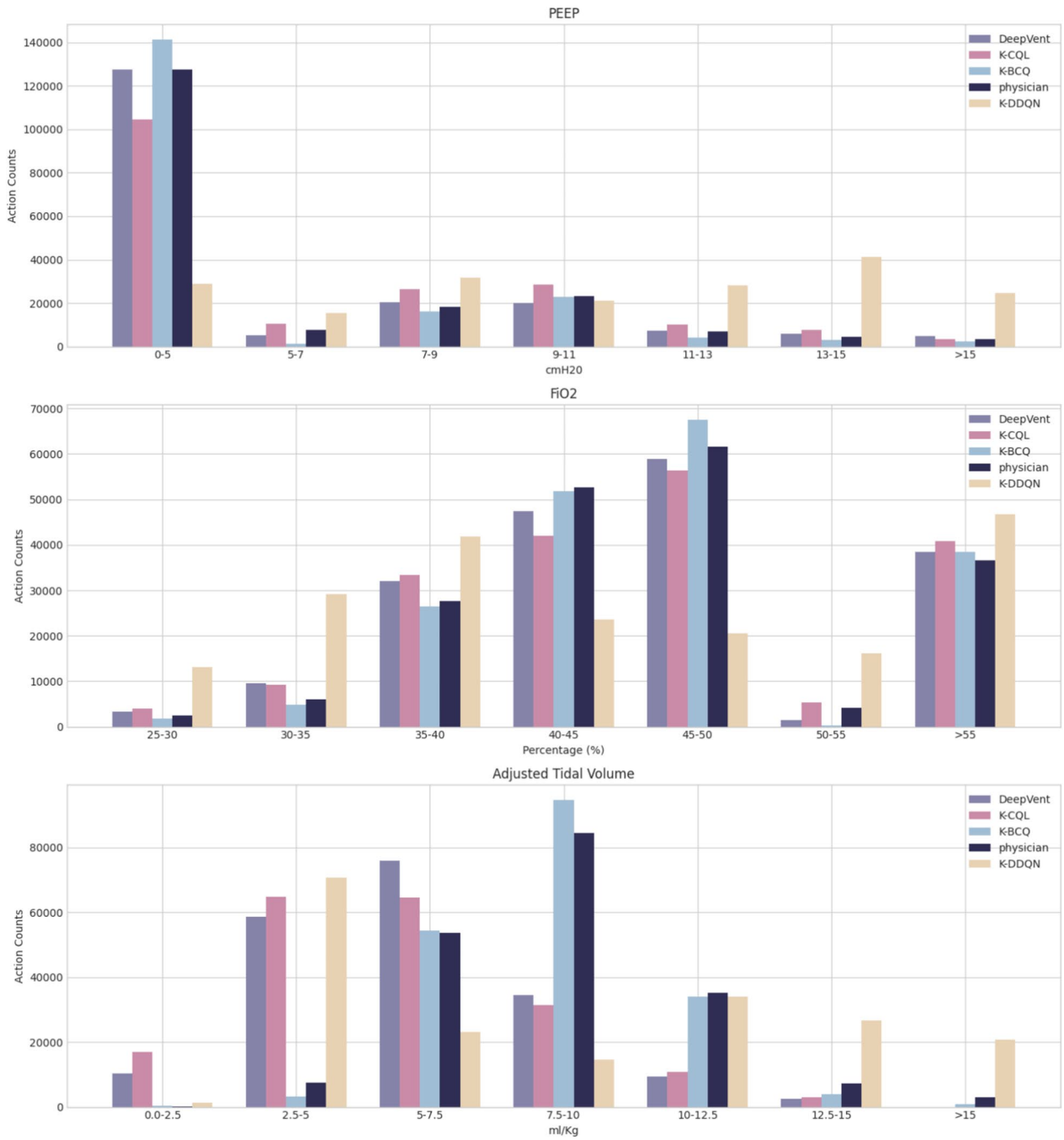

Fig. 4 Ventilator action distribution map of the strategies output by Physician, DeepVent, K-CQL, K-BCQ and K-DDQN. The strategy of K-CQL output meets clinical criteria. The horizontal axis represents the discretized action range, and the vertical axis represents the frequency of each action selected in the test set

intermediate rewards related to ABG analysis further enhances performance. The action distribution plots, used for auxiliary validation, indicate that the K-CQL algorithm has learned to select actions that align with physician preferences, while also considering the long-term consequences of actions to improve therapeutic effect of patient. Moreover, compared to the K-BCQ algorithm, K-CQL exhibits greater stability when confronted with out-of-distribution data, proving the safety of the algorithm and its ability to reduce patient risk.

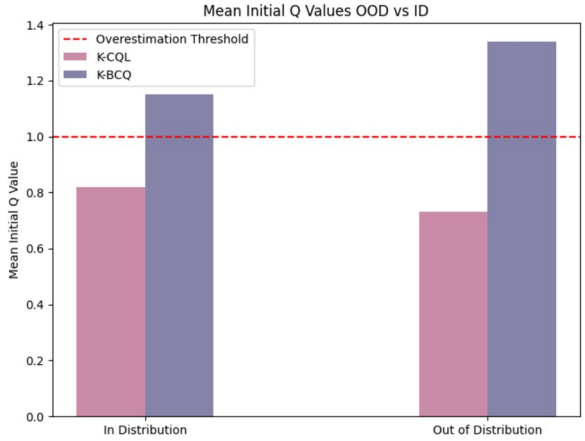

Fig. 5 Initial state estimates of K-BCQ and K-CQL in the state of in-distribution and out-of-distribution. A mean initial q value exceeding 1 is considered q-value overestimation

Funding
This work was supported by Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM (Grant No.2023IHM01083).

Data availability
The MIMIC-III dataset can be obtained from [23].

Declarations

Conflict of interest
The authors have no Conflict of interest to declare that are relevant to the content of this article.

Author details
[1]Department of Automation, University of Science and Technology of China, Shushan District, Hefei 230031, Anhui, China. [2]Institute of Advanced Technology, University of Science and Technology of China, Shushan District, Hefei 230031, Anhui, China. [3]Centre for Leading Medicine and Advanced Technologies of IHM, The First Affiliated Hospital of University of Science and Technology of China, Shushan District, Hefei 230031, Anhui, China.

Received: 11 October 2024 Accepted: 30 July 2025
Published online: 06 August 2025

References

1. Zhang B, Qiu X, Tan X. Balancing therapeutic effect and safety in ventilator parameter recommendation: an offline reinforcement learning approach. Eng Appl Artif Intell. 2024;131: 107784.
2. Zein H, Baratloo A, Negida A, Safari S. Ventilator weaning and spontaneous breathing trials; an educational review. Emergency. 2016;4(2):65.
3. Pham T, Brochard LJ, Slutsky AS. Mechanical ventilation: state of the art. Mayo Clin Proc. 2017. https://doi.org/10.1016/j.mayocp.2017.05.004.
4. Mehedi IM, Shah HSM, Al-Saggaf UM, Mansouri R, Bettayeb M. Adaptive fuzzy sliding mode control of a pressure-controlled artificial ventilator. J Healthcare Eng. 2021;2021:1–10. https://doi.org/10.1155/2021/1926711.
5. Ai W, Suresh V, Roop PS. Development of closed-loop modelling framework for adaptive respiratory pacemakers. Comput Biol Med. 2022;141: 105136. https://doi.org/10.1016/j.compbiomed.2021.105136.
6. Chen Y, Han S, Chen G, Yin J, Wang KN, Cao J. A deep reinforcement learning-based wireless body area network offloading optimization strategy for healthcare services. Health Inf Sci Syst. 2023. https://doi.org/10.1007/s13755-023-00212-3.
7. Yu C, Liu J, Nemati S, Yin G. Reinforcement learning in healthcare: a survey. ACM Comput Surv. 2021;55(1):1–36. https://doi.org/10.1145/3477600.
8. Al-Hamadani MNA, Fadhel MA, Alzubaidi L, Harangi B. Reinforcement learning algorithms and applications in healthcare and robotics: a comprehensive and systematic review. Sensors. 2024;24(8):2461. https://doi.org/10.3390/s24082461.
9. MirzaeeMoghaddamKasmaee A, Ataei A, Moravvej SV, Alizadehsani R, Gorriz JM, Zhang YD, Tan RS, Acharya UR. Elrl-md: a deep learning approach for myocarditis diagnosis using cardiac magnetic resonance images with ensemble and reinforcement learning integration. Physiol Meas. 2024;45(5):055011. https://doi.org/10.1088/1361-6579/ad46e2.
10. Akbulut FP, Akkur E, Akan A, Yarman BS. A decision support system to determine optimal ventilator settings. BMC Med Inform Decis Mak. 2014;14(1):3. https://doi.org/10.1186/1472-6947-14-3.
11. Sutton R, Barto A. Reinforcement learning: an introduction. IEEE Trans Neural Netw. 2005. https://doi.org/10.1109/tnn.2004.842673.
12. Han S, Chen Y, Chen G, Yin J, Wang H, Cao J. Multi-step reinforcement learning-based offloading for vehicle edge computing. In: 2023 15th International Conference on Advanced Computational Intelligence (ICACI), pp 1–8 (2023). IEEE
13. Chen G, Liu X, Shorfuzzaman M, Karime A, Wang Y, Qi Y. Mec-based jamming-aided anti-eavesdropping with deep reinforcement learning for wbans. ACM Trans Internet Technol. 2021;22(3):1–17.
14. Ni J, Huang Z, Cheng J, Gao S. An effective recommendation model based on deep representation learning. Inf Sci. 2021;542:324–42. https://doi.org/10.1016/j.ins.2020.07.038.
15. Chen S, Qiu X, Tan X, Fang Z, Jin Y. A model-based hybrid soft actor-critic deep reinforcement learning algorithm for optimal ventilator settings. Inf Sci. 2022;611:47–64. https://doi.org/10.1016/j.ins.2022.08.028.
16. Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic LB, Schmeink A, Ascheid G, Thiemermann C, Schuppert A, Kindle R, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ Digit Med. 2021;4(1):32.
17. Kondrup F, Jiralerspong T, Lau E, Lara N, Shkrob J, Tran MD, Precup D, Basu S. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 15696–15702 (2023)
18. Yuan Y, Shi J, Yang J, Li C, Cai Y, Tang B. Conservative q-learning for mechanical ventilation treatment using diagnose transformer-encoder. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2346–2351 (2023). IEEE
19. Kumar A, Zhou A, Tucker G, Levine S. Conservative q-learning for offline reinforcement learning. Adv Neural Inf Process Syst. 2020;33:1179–91.
20. Wagner PD. The physiological basis of pulmonary gas exchange: implications for clinical interpretation of arterial blood gases. Eur Respir J. 2015;45(1):227–43.
21. Al Ashry HS, Richards JB, Fisher DF, Sankoff J, Seigel TA, Angotti LB, Wilcox SR. Emergency department blood gas utilization and changes in ventilator settings. Respir Care. 2018;63(1):36–42.
22. Wang Y. Research on key technologies of ventilator intelligent ventilation. Master's thesis, Shandong University; 2023
23. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, Moody B, Szolovits P, AnthonyCeli L, Mark RG. Mimic-iii, a freely accessible critical care database. Sci Data. 2016;3(1):1–9. https://doi.org/10.1038/sdata.2016.35.
24. Suo D, Agarwal N, Xia W, Chen X, Ghai U, Yu A, Gradu P, Singh K, Zhang C, Minasyan E, LaChance J, Zajdel T, Schottdorf M, Cohen D, Hazan E. Machine learning for mechanical ventilation control. Cornell University - arXiv: Cornell University - arXiv; 2021.
25. Siu R, Abbas JJ, Fuller DD, Gomes J, Renaud S, Jung R. Autonomous control of ventilation through closed-loop adaptive respiratory pacing. Sci Rep. 2020. https://doi.org/10.1038/s41598-020-78834-w.
26. Lehman LWH, Adams RP, Mayaud L, Moody GB, Malhotra A, Mark RG, Nemati S. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. IEEE J Biomed Health Inform. 2015;19(3):1068–76. https://doi.org/10.1109/jbhi.2014.2330827.
27. Zhu Y, Zhang J, Wang G, Yao R, Ren C, Chen G, Jin X, Guo J, Liu S, Zheng H, Chen Y, Guo Q, Li L, Du B, Xi X, Li W, Huang H, Li Y, Yu Q. Machine learning

prediction models for mechanically ventilated patients: analyses of the mimic-iii database. Front Med. 2021;8: 662340. https://doi.org/10.3389/fmed.2021.662340.
28. Oruganti Venkata SS, Koenig A, Pidaparti RM. Mechanical ventilator parameter estimation for lung health through machine learning. Bioengineering. 2021;8(5):60. https://doi.org/10.3390/bioengineering8050060.
29. Peng H, Du B, Liu M, Liu M, Ji S, Wang S, Zhang X, He L. Dynamic graph convolutional network for long-term traffic flow prediction with reinforcement learning. Inf Sci. 2021;578:401–16. https://doi.org/10.1016/j.ins.2021.07.007.
30. Ma X, Wang M, Lin S, Zhang Y, Zhang Y, Ouyang W, Liu X. Knowledge and data-driven prediction of organ failure in critical care patients. Health Inf Sci Syst. 2023;11(1):7. https://doi.org/10.1007/s13755-023-00210-5.
31. Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: International Conference on Machine Learning, pp. 2052–2062 (2019). PMLR
32. MacQueen J, et al: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967). Oakland, CA, USA
33. Ahmed M, Seraj R, Islam SMS. The k-means algorithm: a comprehensive survey and performance evaluation. Electronics. 2020;9(8):1295. https://doi.org/10.3390/electronics9081295.
34. Yang W, Bai C, Cai C, Zhao Y, Liu P. Survey on sparse reward in deep reinforcement learning. Comput Sci. 2019;47(03):182–91.
35. Beigmohammadi MT, Amoozadeh L, Rezaei Motlagh F, Rahimi M, Maghsoudloo M, Jafarnejad B, Eslami B, Salehi MR, Zendehdel K. Mortality predictive value of apache ii and sofa scores in covid-19 patients in the intensive care unit. Can Respir J. 2022;2022(1):5129314.
36. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
37. Tang S, Wiens J. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In: Machine Learning for Healthcare Conference, pp. 2–35 (2021). PMLR
38. Nieman GF, Satalin J, Andrews P, Aiash H, Habashi NM, Gatto LA. Personalizing mechanical ventilation according to physiologic parameters to stabilize alveoli and minimize ventilator induced lung injury (vili). Intensive Care Med Exp. 2017. https://doi.org/10.1186/s40635-017-0121-x.
39. Zhou J, Lin Z, Deng XM, Liu B, Zhang Y, Zheng Y, Zheng H, Wang Y, Lai Y, Huang W, Liu X, He W, Xu Y, Li Y, Huang Y, Sang L. Optimal positive end expiratory pressure levels in ventilated patients without acute respiratory distress syndrome: A bayesian network meta-analysis and systematic review of randomized controlled trials. Frontiers of Medicine in China,Frontiers of Medicine in China (2021)
40. Kilickaya O, Gajic O. Initial ventilator settings for critically ill patients. Crit Care. 2013. https://doi.org/10.1186/cc12516.