Article:
Jia, Yan, Kaul, Chaitanya, Lawton, Tom et al. (2 more authors) (2021) Prediction of Weaning from Mechanical Ventilation using Convolutional Neural Networks. Artificial intelligence in medicine. 102087. ISSN: 0933-3657

https://doi.org/10.1016/j.artmed.2021.102087

Prediction of Weaning from Mechanical Ventilation using Convolutional Neural Networks

Yan Jia[a], Chaitanya Kaul[b], Tom Lawton[c], Roderick Murray-Smith[b] and Ibrahim Habli[a]

[a]Department of Computer Science, University of York, York, UK
[b]School of Computing Science, University of Glasgow, Glasgow, UK
[c]Bradford Royal Infirmary and Bradford Institute for Health Research, Bradford, UK

ARTICLE INFO

Keywords:
Ventilator weaning
Mechanical ventilation
Deep learning
Feature importance

ABSTRACT

Weaning from mechanical ventilation covers the process of liberating the patient from mechanical support and removing the associated endotracheal tube. The management of weaning from mechanical ventilation comprises a significant proportion of the care of critically ill intubated patients in Intensive Care Units (ICUs). Both prolonged dependence on mechanical ventilation and premature extubation expose patients to an increased risk of complications and increased health care costs. This work aims to develop a decision support model using routinely-recorded patient information to predict extubation readiness. In order to do so, we have deployed Convolutional Neural Networks (CNN) to predict the most appropriate treatment action in the next hour for a given patient state, using historical ICU data extracted from MIMIC-III. The model achieved 86% accuracy and 0.94 area under the receiver operating characteristic curve (AUC-ROC). We also performed feature importance analysis for the CNN model and interpreted these features using the DeepLIFT method. The results of the feature importance assessment show that the CNN model makes predictions using clinically meaningful and appropriate features. Finally, we implemented counterfactual explanations for the CNN model. This can help clinicians understand what feature changes for a particular patient would lead to a desirable outcome, i.e. readiness to extubate.

1. Introduction

Mechanical ventilation via an endotracheal tube, sometimes also called invasive mechanical ventilation, is one of the most widely used interventions for patients admitted to intensive care units (ICUs). Mechanical ventilation is a life-saving medical procedure used to assist or replace spontaneous breathing for patients with acute respiratory difficulties. Studies have shown that around 40% of ICU patients require invasive mechanical ventilation [1]. This consumes significant ICU resources with estimated daily costs around £1,738 in the UK [2] and $2,300 in the US [3].

Weaning patients from mechanical ventilation covers the process of liberating the patient from mechanical support and removing the endotracheal tube (extubation). Time spent in this weaning process occupies a significant proportion of the total duration of mechanical ventilation [4]. Assessment of weaning readiness is a complex clinical task, which often includes determining whether or not the underlying disease of the patient has been successfully treated, together with haemodynamic stability, the patient's level of consciousness, and the current values for ventilator settings. The final stage is often to conduct a series of Spontaneous Breathing Trials (SBTs), using either unsupported T-piece breathing or low-level Pressure Support Ventilation (PSV) over at least 30 minutes [5].

Despite advances in medical knowledge, weaning too early or too late are still problematic. Delays in assessing readiness to wean are a common cause of late weaning. As a consequence, patients with prolonged ventilation might experience airway trauma, post-extubation delirium, drug dependencies, ventilator induced pneumonia, other forms of increased morbidity and even higher fatality rates [6] [7] [8]. There are also non-clinical effects including increased costs and greater strain on hospital resources, e.g. it has been reported that patients on prolonged ventilation use 37% of ICU resources [9].

On the other hand, premature extubations may lead to extubation failure, where re-intubation is required within 48-72 hours. Studies have shown that up to 25% of patients suffer extubation failure due to recurrence of respiratory

⋆This document is the result of a research project funded by Bradford Teaching Hospitals NHS Foundation Trust.
⋆⋆Chaitanya Kaul and Roderick Murray-Smith acknowledge support from the iCAIRD project, funded by Innovate UK (project number 104690).
ORCID(s): 0000-0002-5446-6565 (Y. Jia); 0000-0003-4893-6222 (C. Kaul); 0000-0003-1351-8127 (T. Lawton); 0000-0003-4228-7962 (R. Murray-Smith); 0000-0003-2736-8238 (I. Habli)

insufficiency and require re-intubation [10], which can cause severe patient discomfort and result in even longer stays in the ICU with associated increases in cost and resource demands [11]. As with early extubation there can be increased fatality rates [12].

Considering the risks of prolonged dependence on mechanical ventilation and premature extubation, it is important to identify the ideal time point for weaning from mechanical ventilation from both a patient and healthcare provider point of view. However, there is no consensus on a standardised weaning protocol [13], even though they can be of benefit [14]. In practice protocols can vary between institutions, and may include different parameters [15]. This is mainly due to uncertainty, so an automated prediction model to indicate when extubation may be appropriate is likely to be helpful to clinicians seeking to make better-informed decisions.

In this paper we report a decision support model that aims to advise clinicians when a patient is ready for extubation, using patient information which is routinely available in the ICU setting. More specifically, we employ Convolutional Neural Networks (CNNs) to predict extubation readiness at each patient state, with a one hour time step. The model was developed using the MIMIC-III clinical database [16] and incorporates 25 patient features, such as demographics, vital signs and laboratory values. Our goal is to help the clinicians choose the most appropriate action at each time step, including initiation of an SBT or commencing extubation, and to provide explanations that give insights into the model predictions and hence improve confidence to act on those predictions.

This paper describes three main components to a system aiming to realise this goal; they are summarised as follows:

- CNN to predict the most appropriate treatment action in the next hour for a given patient state, with 86% accuracy and 0.94 AUC-ROC performance;
- feature importance assessment showing that the CNN model is making prediction using clinically meaningful features; and
- counterfactual explanations that help clinicians understand what kinds of feature changes for a particular patient would lead to a desirable outcome, i.e. readiness to extubate.

The paper is organised as follows: Section 2 describes related work, focusing on the use of machine learning (ML) for helping clinicians with weaning decisions. In Section 3, we describe the data and methods used, and Section 4 presents the results, i.e. the performance of the CNN, analysis of feature importance and counterfactual explanations. Our results are discussed in Section 5, and we set out our conclusions in Section 6.

2. Related Work

The growing use of electronic health records (EHRs) has enabled data-driven approaches to healthcare, including employing ML for diagnosis and to recommend treatments. A 2017 survey found over 5,000 publications with the majority using support vector machines and neural networks, although more than ten different ML methods have been reported [17]. The use of ML methods offers the prospect of more effective healthcare, including personalised treatment regimes. In this section, we focus on ventilator weaning and consider the different ML methods and sets of features included in the models that have been explored in this clinical context.

Supervised learning methods, e.g. statistical methods such as logistic regression, Artificial Neural Networks (ANNs), or naive Bayes have promise for predicting extubation outcome for infants [18]. ANNs have been shown to give better results than standard clinical criteria such as the Rapid Shallow Breathing Index (RSBI) and maximum inspiratory pressure [19]. An association rule network-based feature category-weighted naive Bayes method has been proposed to better support physicians' weaning decision-making, and was shown to consistently outperform other benchmark techniques, e.g. support vector machine (SVM) [20]. Other recent work has used reinforcement learning to develop policies for ventilation weaning [21]. There is no apparent consensus on the best approach to use, although ANNs are widespread, e.g. [19] [22].

A wide range of features have been considered for predicting extubation failure. These include demographic information (e.g., age, reason for intubation) [19], vital signs (e.g., heart rate, respiratory rate) [23], blood gas analysis (e.g., sodium, potassium, serum anion gap, oxygen/carbon dioxide partial pressure) [24], and respiratory parameters (e.g., duration of mechanical ventilation, tidal volume) [19] [22]. What is striking is the variation in the factors used in different studies:

- subjects' age, reasons for intubation, duration of mechanical ventilation, Acute Physiology And Chronic Health Evaluation (APACHE II) scores, and breathing patterns obtained during a 30-minute SBT [19];

- tidal volume, minute ventilation, breathing frequency, and maximum inspiratory pressure [22];
- pre-extubation serum anion gap values and ratio of arterial oxygen partial pressure to fractional inspired oxygen (P:F ratio) [24];
- signal power of the respiratory flow obtained during the inspiratory phase [25];
- cardiorespiratory behaviour [26] and respiratory pattern parameters [27].

Generally, these studies use small numbers of features in training and prediction. In contrast, the reinforcement learning approach mentioned earlier uses 32 features [21]. Some work reports systematic approaches to identifying the relevant features. One study has sought to identify relevant features by comparing all combinations of three features from a total of 57,[23] eventually selecting 6 features from the best two models. Work using a Light Gradient Boosting Machine [28] initially considers 92 features and reduces them to 36 for the final model. This work [28] and the reinforcement learning approach [21] analyse feature importance to help to interpret the models.

Our approach is different from, and complements, the existing work. Most of the previous work predicts extubation outcomes whereas we monitor patient states every hour and indicate when patients are potential candidates for extubation thus prompting clinicians to commence a SBT. In ML terms our work employs deep learning, specifically CNNs, with a comparatively rich model (25 features) giving better performance, and uses counterfactual explanations as well as feature importance to help clinicians have confidence in the system's predictions. We see our approach as a step towards personalised and actionable healthcare.

3. Methods

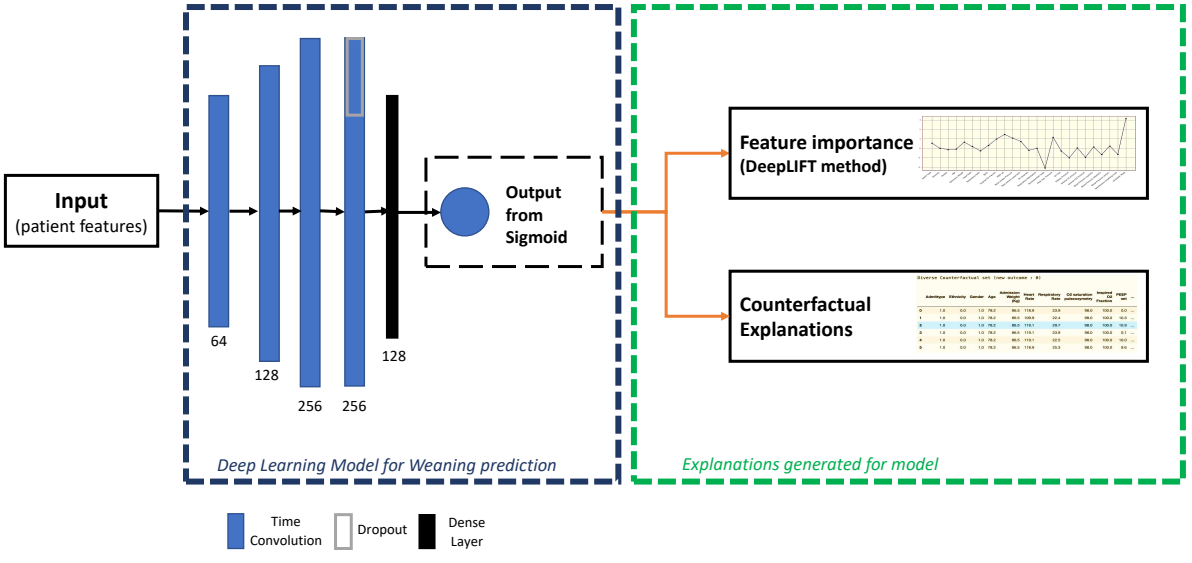

Figure 1: Overview of the method for the paper

We developed an end-to-end actionable and explainable deep learning prediction system that can assist clinicians in making decisions about weaning from mechanical ventilation. As seen from Figure 1, our approach is broadly divided into two connected parts – the model itself, and the mechanisms to generate explanations for the model. The input data, comprising of 25 patient features, is passed into a series of convolution layers with a varying number of

filters to extract relevant features, which are then passed into the output layer that provides a prediction readiness to extubate within the next one hour. The explanation for this prediction is then generated in two different ways: feature importance and counterfactual explanations. For a complex ICU setting, it is important to create a system that not only produces predictions, but also provides different explanations along with it. To this end, we use DeepLIFT [29] to generate per patient feature importance graphs. We also generate multiple diverse counterfactual examples using DiCE [30], which can help clinicians to identify and choose what kind of features to change to achieve the desired outcome.

3.1. Critical Care Data

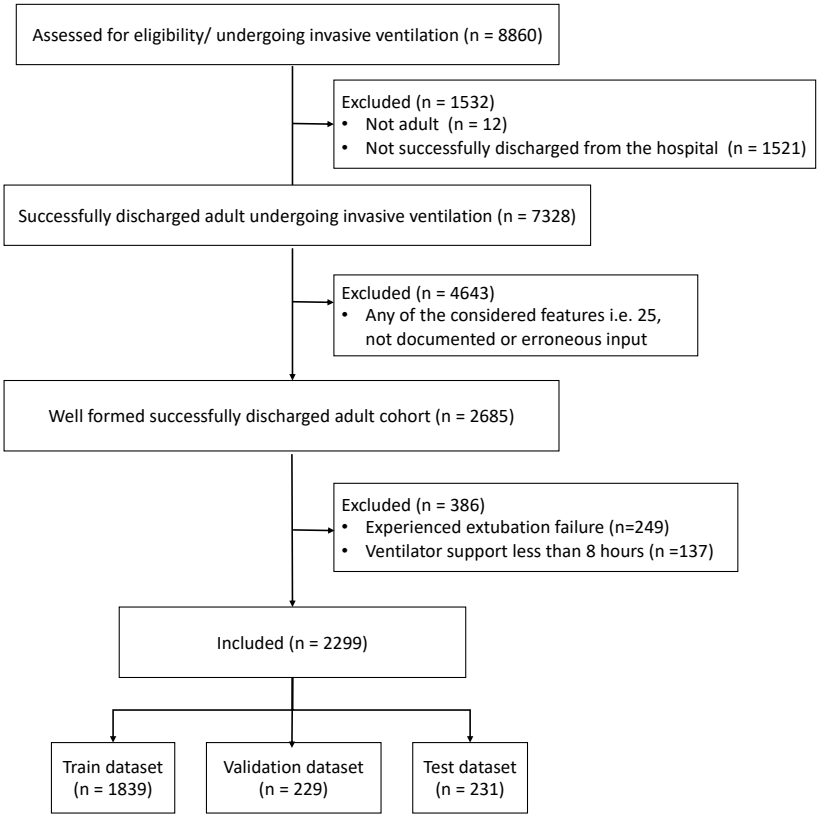

Figure 2: Patient inclusion diagrams in MIMIC-III

We used Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC-III) [16] a freely available data set to develop the model. The MIMIC-III data set was collected from the Beth Israel Deaconess Medical Center in Boston and contains patient demographics, vital signs, records of fluid and medication administration, results of laboratory tests, observations and notes provided by care professionals. We initially selected 8,860 admissions who underwent invasive mechanical ventilation from the data set. We excluded non-adult patients and also those who died in the hospital as these fatalities can be caused by factors that are beyond the weaning process, in line with other work [21] [28]. This resulted in 7,328 adult patients who were successfully discharged following invasive ventilation, see Figure 2.

Based on the literature surveyed, e.g. clinical studies of "protocolized" weaning [31], and clinical judgement, we extracted 25 features including patient demographics, e.g. age, gender, ethnicity, laboratory tests, e.g. arterial pH, and vital signs, e.g. heart rate, oxygen saturation (SpO_2), and ventilator information, e.g. ventilator mode, Positive End-Expiratory Pressure (PEEP), and mean airway pressure. We took the patient data and produced a series of records with values for the features on an hourly basis, for each patient from when the ventilator mode was recorded until the last time it was recorded. This ensured that the records covered the whole invasive ventilation period and also the non-

invasive ventilation support period (n.b. non-invasive modes were also included). Where multiple values for a feature were available in an hour, they were averaged. Further, data in MIMIC-III is not available for every hour so where individual values of a feature were missing they had to be estimated. We used the previous valid value, if available, to fill in the missing values i.e. forward propagation; if there was no valid previous value then back propagation was used. After this process, if a patient record still had missing values for some features, i.e. no values were recorded during the period considered, or was obviously erroneous then they were deleted. This processing resulted in a well-formed, successfully discharged adult cohort of 2,685 patients, see Figure 2. The main reason for the substantial reduction in number of patients is the absence of values of some features, as deep learning will not accept missing data during training.

As a final processing stage we excluded patients who had ventilation support for less than 8 hours as they were likely to be undergoing routine ventilation following elective surgery. Post-operative extubation is minimal risk of adverse extubation outcomes and it was not our intention to consider such cases in this work. Patients who experienced extubation failure were also excluded to ensure the predictions relate to successful weaning, because extubation failure could arise from premature extubation [11] [32], and thus the weaning schedule for such patients could be misleading. To be consistent with previous studies we defined extubation failure as the need for re-intubation within 48 hours [5] [33]. This produced a final cohort of 2,299 patient admissions for use in our study, see Figure 2.

3.2. Convolutional Neural Networks

CNNs have proven useful in image analysis, and their application has been explored in various other domains such as time series forecasting and data generation. The rationale for using a CNN for this task is that they are fast at run time and have the potential to produce accurate predictions for the type of tabular data employed here, see for example [34] [35] [36] [37].

CNNs make predictions by extracting features without explicit, pre-defined knowledge of what is important in the data. In CNNs convolution computations are generally followed by non-linearities, also known as activation functions. The most commonly used activation is the Rectified Linear Unit (ReLU), given by $ReLU(x) = \max(0, x)$, where the response of a network is zeroed for negative values of the features learnt. Stacking multiple layers of convolutions and activation functions together extracts features in a CNN. These features are then passed into fully connected layers that learn to make the prediction.

The architecture of our CNN went through extensive tuning. The input features are passed through a series of 4 convolution layers with filter sizes 64, 128, 256, 256 and dropout is used in the final convolution layer. The output from this convolution layer is then flattened and passed into a fully connected layer of size 128 nodes which is then fed into the output layer, making the prediction through a sigmoid function with a threshold set at 0.5. The architecture of the CNN model is summarised in Table 1 and in Figure 1.

Table 1

CNN architecture
Conv1D with 64 filters of Kernel size 1
Conv1D with 128 filters of Kernel size 1
Conv1D with 256 filters of Kernel size 1
Conv1D with 256 filters of Kernel size 1
Dropout with probability 0.5 of leaving out units
Fully connected layer with 128 neurons
Sigmoid output

We use the following metrics to assess our work: accuracy, precision, recall, F1 score, as well as the area under the ROC (AUC-ROC) curve. This is the generally accepted set of evaluation metrics for deep learning.

We calculate the accuracy of our model as the ratio of the number of correct predictions to the total number of predictions. Formally:

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

where, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives in the predictions.

Similarly, precision is given by:

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

and can be interpreted as the proportion of the positive predictions that were correct. Recall is given by:

$$Recall = \frac{TP}{TP + FN}$$

and can be interpreted as the true positive rate (i.e. the number of true positives divided by the total number of elements that actually belong to the positive class). A model can have a high precision or recall and do badly on the other metric. An F1 score takes both scores into account so as to better evaluate the model's performance. It is given by:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The ROC curve demonstrates the model's ability to provide predictions at various decision thresholds. It assesses how well a model can distinguish between the classes and is a plot of the True Positive Rate (TPR) of a model against its False Positive Rate (FPR). The area under the curve (AUC) denotes the probability of the classifier ranking a random positive sample in the data higher than a random negative sample.

3.3. Explainability

ML systems are often "opaque" in that it is not easy for users to understand the reasoning behind the prediction [38]; for this reason they are sometimes referred to as "black box" [39]. The term "explainability" is used for methods that help stakeholders, e.g. users or developers, understand why the ML models behave in a certain way. Explainability is particularly important in healthcare where decisions made with the support of ML models can have an impact on patient safety [40].

There are a range of methods for achieving explainability. Here we consider two relevant, and complementary, classes of method: feature importance and counterfactual explanations, and present the specific methods that we have used.

3.3.1. Feature importance

Feature importance methods can help clinicians gain insights behind ML model predictions. They involve identifying the features in a model that are most significant in making a prediction; more specifically, the features are normally ranked in order of importance. Feature importance, sometimes called feature attribution, is by far the most common explainability method [41] [42].

Generally, feature importance methods for complex ML models try to build a simpler model than the original one (sometimes known as the "explanation model"), as the original model is hard to interpret. Lundberg has pointed out that many current feature importance methods use the same explanation model, which is a linear function summing the effects of all feature attributions to approximate the output of the original model; methods that match this definition are called additive feature attribution methods [43].

In this paper we use a feature importance method known as DeepLIFT (Deep Learning Important FeaTures) [29] which is one of these additive feature attribution methods. It has been developed specifically for use with deep NNs. When explaining deep NNs, the features are the set of inputs to the model. DeepLIFT compares the activation of each neuron to its "reference activation" and attributes to each input an importance score according to the difference. The "reference activation" is obtained through some user-defined reference input to represent an uninformative background value, for example for image classification this could be a totally black image. Research shows that DeepLIFT can be viewed as a variant of gradient-based methods where the gradient for the non-linearity is calculated using the ratio between the difference in output and the difference in input and the gradient for the linearity is just the weights [44]. We chose this method for three main reasons. First, DeepLIFT considers both positive and negative contributions of features, thus exposing the sign of dependencies from the input features to the output. Second, it deals effectively with discontinuities in the gradient of the CNN model as it uses a difference from reference approach. Third, it avoids the problem of model saturation where using gradients would just assign zero to the features. DeepLIFT can still assign a non-zero score to the features. In addition, the feature ranking is generated by a single backpropagation through the network so the explanation can be generated efficiently, which could help the clinician by providing the explanation at the time of making a prediction.

3.3.2. Counterfactual explanations

Post-hoc explanations can enable clinicians to understand predictions of ML models and decide how to act to achieve the desired outcomes. Counterfactual explanations for ML, which were introduced by Wachter et al [45], are one way to achieve this. Counterfactual explanations are proposed as a way to provide a different and more desirable outcome by perturbing the input of the ML model. For example, when the CNN model predicts that the patient should continue with mechanical ventilation, counterfactual explanations would provide the clinician with information on what features of this patient need to change, such as successful completion of an SBT, in order to change the prediction. To be useful, counterfactual explanations should minimise the difference (distance) from the current inputs to the counterfactual examples. Further, it can be helpful to produce diverse counterfactual explanations to give clinicians a choice of what features to change, given the feasibility of the change, to achieve the desired outcome.

Generating counterfactual explanations can be viewed as an optimisation problem: given an input x, a model f, and a distance metric d, we find a counterfactual explanation c as follows:

$min\ d(x, c)$

such that $f(c)$ provides a desirable outcome. DiCE (Diverse Counterfactual Explanations) is the method we used here [30], which can produce diverse counterfactual examples to help the clinicians to choose features that it is feasible to change. We use this approach as it enables us to provide counterfactual explanations that are tailored to individual patients and are actionable.

4. Results

4.1. Performance of the CNN

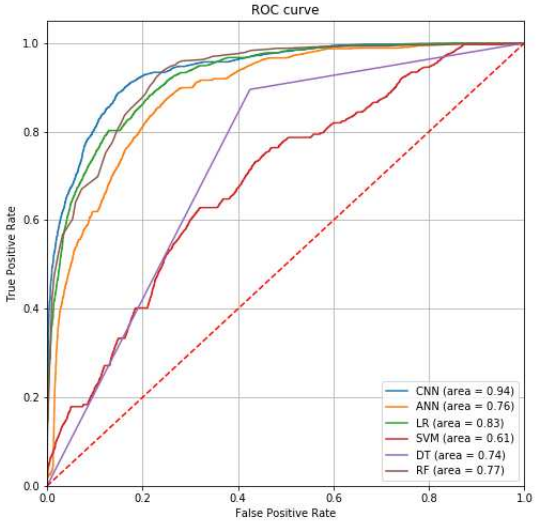

Figure 3: CNN Performance when trained on 1,839 records from MIMIC-III

As shown in Figure 2, the included patient admissions were split into three sets: 80% (n = 1,839) were randomly assigned to the training set; 10% (n = 229) to the validation set for tuning the hyper-parameters of the model; and 10% (n = 231) to the test set for evaluating the final performance of the model. To prevent over-fitting during model training, L2 regularisers were used and a comparison of the performance between the training and validation data sets was undertaken. There was no significant difference in accuracy between the validation set (86%) and training set (87%) regarding the CNN model.

The AUC-ROC using the test set for the CNN model was 0.94 which was better than any of the following ML methods: 0.76 for ANN, 0.83 for Logistic Regression, 0.61 for Support Vector Machine, 0.74 for Decision Tree and 0.77 for Random Forest Tree, as shown in Figure 3.

Predicting extubation readiness using the CNN model on the test set has a precision of 82%, a recall of 86% and an accuracy rate of 86% with the optimal threshold of 0.5. The accuracy, precision and recall for predicting extubation readiness from the test set using other ML methods are listed in Table 2. This shows that the CNN model has generally superior performance to any of the other methods presented here. All these methods show good results in accuracy and precision, but they have comparatively low recall and AUC-ROC.

Table 2
Performance comparison with different ML classifiers

Methods	Accuracy	Precision	Recall	F1-Score	AUC
CNN	86%	82%	86%	84%	0.94
ANN	85%	84%	76%	79%	0.76
Logistic Regression	82%	78%	84%	79%	0.83
Support Vector Machine	70%	61%	61%	61%	0.61
Decision Tree	81%	76%	74%	74%	0.74
Random Forest Tree	87%	90%	77%	80%	0.77

4.2. Feature importance analysis

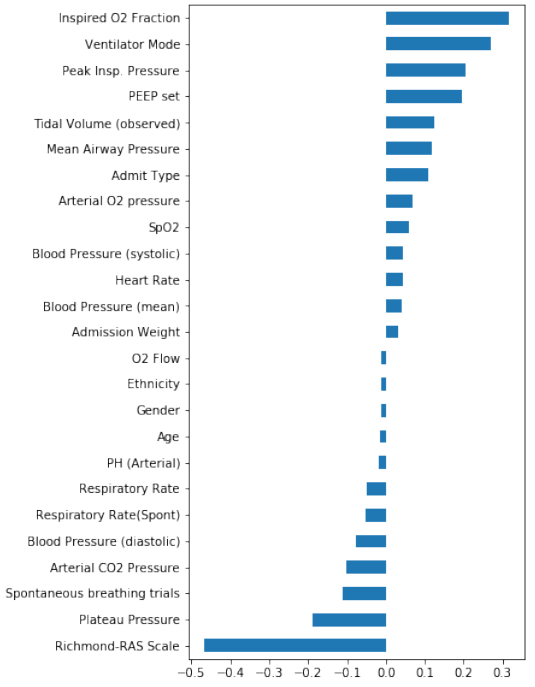

Figure 4: Feature Importance for the CNN Model

An overview of the feature importance produced using DeepLIFT is shown in Figure 4. In this work, the reference sample is the minimum values of all of the input features obtained from the data set, for which the CNN gives a prediction of 0.13. A positive feature importance score contributes to moving the output above its reference value towards a patient remaining intubated. In contrast, a negative feature importance score contributes to moving the output below the reference value, which supports a prediction of extubation. Those features that score near zero, e.g. ethnicity, gender and age, are unimportant and have little influence on the prediction. This shows the model does not rely on sensitive attributes that can inadvertently lead to bias.

Other aspects of the ranking correlate well with clinical expectations, helping to give confidence in the model. Patients who are undergoing invasive mechanical ventilation are often sedated to maintain physiological stability and

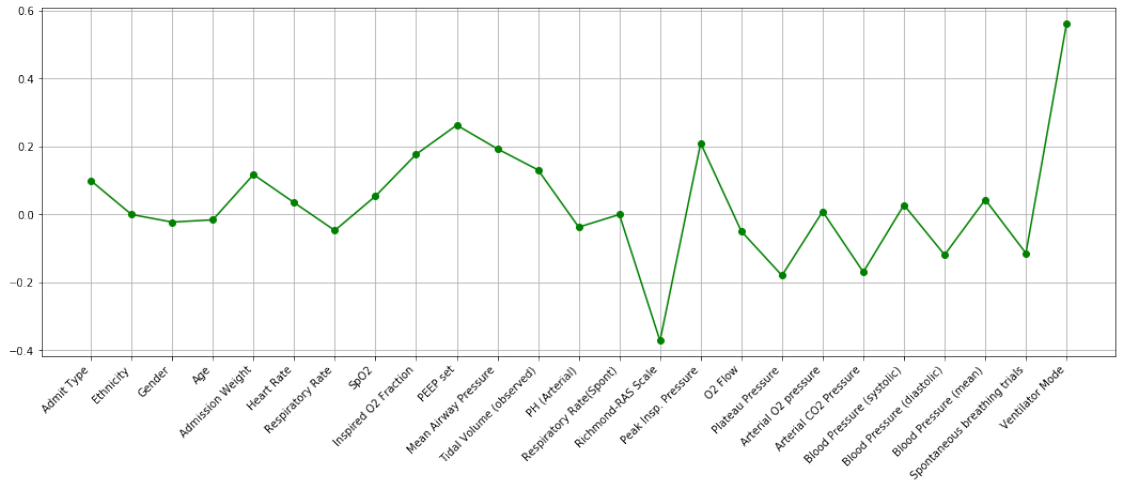

Figure 5: Feature Importance for a single patient

to control pain levels. Sedation is reflected in the Richmond Agitation Scale (RAS) with negative values representing sedation and 0 meaning that they are alert and calm, thus more likely to be suitable for weaning. This is of particular practical importance as sedation is under the clinician's direct control. Also, a positive SBT result contributes to confidence that the patient is ready to extubate. The factors with a strong positive weight, e.g. inspired O_2 fraction, ventilator category, peak inspiratory pressure, and PEEP set, are also important clinical factors that a doctor will consider in the weaning decision-making process.

Figure 5 visualises the feature importance values for a *single* patient; positive and negative scores for features are interpreted the same way as in Figure 4. The sum of the positive contributed features are far greater than the sum of the negative contributed features, thus the prediction for this patient, for the next hour, is that they will remain intubated.

4.3. Counterfactual explanations

Table 3 shows a single counterfactual example for the patient shown in Figure 5, to illustrate the method. Certain features cannot be varied, e.g. age and gender; the dash in the right hand column indicates no change from the original input. The figures in the right hand column show feature changes that would achieve a different prediction, as shown in the bottom row. This example could potentially help the clinician to identify actions to take so that the patient becomes ready for extubation. It is also possible to generate multiple counterfactual examples, so that the clinicians can choose one that is most practical to implement.

5. Discussion

With the wide use of EHR systems, it becomes possible to feed healthcare data from EHR to ML-based decision support systems (DSS). Research in recent years has shown the potential benefit from using ML in healthcare. Our work on decision support for weaning is but one example. However, in order to successfully deploy such systems in healthcare, there are many factors to consider. Demonstrating a good performance of the ML model is a minimum criterion for accepting this kind of system. Given that clinicians are already overloaded, it is also important to ensure that such systems decrease the burden on clinicians. For example, a user friendly interface is as important as the ML model itself, no matter whether it is integrated with the EHR or a stand-alone interface. To achieve good ML model performance, especially when using deep learning, a lot of patient features are often required (in our case 25 features). Thus, it would be desirable to integrate the ML model with the EHR system rather than rely on manual input, avoiding further error from transcribing. Furthermore, safety has also been expressed as one of the major concerns for such systems, especially in critical areas of healthcare [46] [47], and this is an active area of study by regulators, e.g. [48] [49]. Explainability can also contribute to safety, especially in healthcare which is a complex, social and technical organisation. Thus, explanations of what data is required and how it is pre-processed are also important, and this can ensure that the model is used correctly. For example, some features in our model are frequently monitored, e.g. SpO_2,

Table 3
An Example of generated counterfactual for the specific patient in Figure 4

Features	Original input	Counterfactual Example
Admit Type	Emergency	—
Ethnicity	White	—
Gender	Female	—
Age	78.2	—
Admission Weight	86.5	—
Heart Rate	119	110
Respiratory Rate	24	—
SpO_2	98%	—
Inspired O_2 Fraction	100%	40%
PEEP set	10	5
Mean Airway Pressure	14	10
Tidal Volume (observed)	541	—
PH (Arterial)	7.46	—
Respiratory Rate(Spont)	0	24
Richmond-RAS Scale	-1	0
Peak Insp. Pressure	21	—
O_2 Flow	5	—
Plateau Pressure	19	—
Arterial O_2 pressure	124	118
Arterial CO_2 Pressure	33	—
Blood Pressure (systolic)	101	—
Blood Pressure (diastolic)	65	—
Blood Pressure (mean)	76	—
Spontaneous breathing trials	No result	Successfully Completed
Ventilator Mode	CMV/ASSIST/AutoFlow	SIMV/PSV
Predicted outcome	0.93	0.17

but other features are less so, e.g. arterial blood gas. Explaining how features which are not monitored hourly are pre-processed is also important - in our case, the value for such features are forward propagated if there is no current value. In addition, explaining the ML model itself is also important as a way of gaining confidence in such systems. In this paper, we present two approaches to explainability, which we consider to be complementary. For the feature importance method, the overview (see Figure 4) gives clinicians confidence that the model is considering important clinical factors, e.g. ventilator mode and peak inspiratory pressure, rather than fixed attributes of the patient, e.g. ethnicity and gender. The individual patient feature importance (see Figure 5) gives clinicians information that helps them place trust in predictions for a specific patient. The counterfactual explanations also focus on the specific patient helping clinicians to see what actions to take to increase the chances of a successful extubation.

However, there is a relationship between the DeepLIFT method and counterfactual explanations. DeepLIFT can be viewed as a variant of gradient-based methods where the gradient for non-linearities is modified (see Section 3.3.1). The importance score given by DeepLIFT is equal to $(x - x')$ multiplied by the modified gradient, where x is the input features and x' is the "reference activation", see the proof [44]. The input features for our CNN model are normalised between 0 and 1, so the minimum value of the input features is a zero-array after normalisation. Thus, in this case the importance is defined as $x \times$ the modified gradient. Where the feature has a higher score using DeepLIFT, i.e. the absolute score for a feature is greater than zero, perturbation of this feature will make a larger difference in the prediction given that the input feature values are on a similar scale. As counterfactual explanations are proposed as a way to provide perturbations that would have changed the prediction of a model, we expect a correlation between the importance scores produced by DeepLIFT and the counterfactual explanations. This is observed in our experiments. In the counterfactual example, the change of ventilator mode, RAS scale, PEEP set and Inspired O_2 fraction produce a different prediction, where these features are shown to have high importance score in Figure 4. Therefore, changes to these features will help to "flip" the prediction.

It is also worth noting that Respiratory rate (spontaneous) has a zero score as shown in Figure 5. This is because

the input feature value is zero. Therefore, in this case, this feature is not important but this might not be true in other cases. In the counterfactual example, arterial O_2 pressure also changed from 124 to 118, even though this feature only weighs 0.05 in Figure 5. If we investigate the counterfactual example by changing the arterial O_2 pressure back to 124, the new prediction is 0.169, which is a small change from 0.168 (this is the original prediction of the counterfactual example in Table 3). This is consistent with our expectation as it has a small importance score in Figure 5.

The work also has its limitations. Most importantly, our model only incorporates 25 features. Whilst this is far more than in many other approaches, we have no evidence to show that this is optimal and that benefits would not accrue from the inclusion of additional features, as has been tried elsewhere, e.g. [26] [28]. For example, some work has investigated the use of diaphragmatic parameters to predict the outcome of weaning from mechanical ventilation and found that time to peak inspiratory amplitude of the diaphragm exhibits good performance in predicting the success of weaning [50]. It would be interesting to see the effect of including diaphragmatic parameters in the ML model, but unfortunately such features are not recorded in MIMIC III. In addition, CT scans or chest X-rays are also factors that doctors will consider when they are making weaning decisions. However, such diagnostic tests are images supported with free text radiology reports. Current ML methods are limited in their ability to integrate different types of data (e.g. tabular, text and images) in a single model. However, if we can find a way to integrate these different forms of data into one model, it might enable us to improve performance. This is a potential area for future work. However, we would continue our emphasis on explainability as well as comparing model performance. If, for example, additional features gained a low weighting in overall feature importance, even though performance increases slightly, then this might suggest that they are not worth including in the model.

6. Conclusion

In this paper, we have employed a CNN to predict extubation readiness from mechanical ventilation, and achieved promising performance. We have also explored the explainability of the CNN model using both feature importance and counterfactual explanations, and showed that they can complement each other. Feature importance gives general confidence in the learnt model, whereas the counterfactual explanations help clinicians see what actions might improve a patient's chances of a successful extubation. We believe that, used in this way, explainability is a step towards making ML-based DSS both personalised and actionable. We also see these explainability methods as contributing to the safety of the DSS, with the potential to influence standards for safety-related systems such as clinical DSS.

The code for data processing and learning the CNN model along with the code for generating the explanations is available at: https://github.com/Yanjiayork/mechanical_ventilator.

Acknowledgement

This work is funded by Bradford Teaching Hospitals NHS Foundation Trust and supported by the Assuring Autonomy International Programme at the University of York. We also acknowledge support from the iCAIRD project, funded by Innovate UK (project number 104690). We thank Dr Niranjani Prasad from Princeton University for sharing part of code for extracting data from MIMIC-III. The views expressed in this paper are those of the authors and not necessarily those of the NHS, or the Department of Health and Social Care.

References

[1] H. Wunsch, J. Wagner, M. Herlim, D. Chong, A. Kramer, and S. D. Halpern, "Icu occupancy and mechanical ventilator use in the united states," *Critical care medicine*, vol. 41, no. 12, 2013.

[2] J. Marti, P. Hall, P. Hamilton, S. Lamb, C. McCabe, R. Lall, J. Darbyshire, D. Young, and C. Hulme, "One-year resource utilisation, costs and quality of life in patients with acute respiratory distress syndrome (ards): secondary analysis of a randomised controlled trial," *Journal of intensive care*, vol. 4, no. 1, p. 56, 2016.

[3] L. M. Cooper and W. T. Linde-Zwirble, "Medicare intensive care unit use: analysis of incidence, cost, and payment," *Read Online: Critical Care Medicine| Society of Critical Care Medicine*, vol. 32, no. 11, pp. 2247–2253, 2004.

[4] T. Chockalingam, "Weaning and extubation," *J Lung Pulm Respir Res*, vol. 2, no. 3, p. 00043, 2015.

[5] J.-M. Boles, J. Bion, A. Connors, M. Herridge, B. Marsh, C. Melot, R. Pearl, H. Silverman, M. Stanchina, A. Vieillard-Baron, *et al.*, "Weaning from mechanical ventilation," *European Respiratory Journal*, vol. 29, no. 5, pp. 1033–1056, 2007.

[6] L. M. Bigatello, H. T. Stelfox, L. Berra, U. Schmidt, and E. M. Gettings, "Outcome of patients undergoing prolonged mechanical ventilation after critical illness," *Critical care medicine*, vol. 35, no. 11, pp. 2491–2497, 2007.

[7] C. G. Hughes, S. McGrane, and P. P. Pandharipande, "Sedation in the intensive care setting," *Clinical pharmacology: advances and applications*, vol. 4, p. 53, 2012.

[8] A. Esteban, A. Anzueto, F. Frutos, I. Alía, L. Brochard, T. E. Stewart, S. Benito, S. K. Epstein, C. Apezteguía, P. Nightingale, *et al.*, "Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study," *Jama*, vol. 287, no. 3, pp. 345–355, 2002.

[9] D. Wagner, "Economics of prolonged mechanical ventilation," *American Journal of Respiratory and Critical Care Medicine*, vol. 140, 1989.

[10] M. J. Tobin, "Advances in mechanical ventilation," *New England Journal of Medicine*, vol. 344, no. 26, pp. 1986–1996, 2001.

[11] J. S. Krinsley, P. K. Reddy, and A. Iqbal, "What is the optimal rate of failed extubation?," *Critical Care*, vol. 16, no. 1, pp. 1–5, 2012.

[12] J. Whiting, J. Gowardman, D. Huntington, *et al.*, "The effect of extubation failure on outcome in a multidisciplinary australian intensive care unit," *Critical Care and Resuscitation*, vol. 8, no. 4, p. 328, 2006.

[13] G. Conti, J. Mantz, D. Longrois, and P. Tonner, "Sedation and weaning from mechanical ventilation: time for 'best practice'to catch up with new realities?," *Multidisciplinary respiratory medicine*, vol. 9, no. 1, p. 45, 2014.

[14] H. M. Horst, D. Mouro, R. A. Hall-Jenssens, and N. Pamukov, "Decrease in ventilation time with a standardized weaning process," *Archives of Surgery*, vol. 133, no. 5, pp. 483–489, 1998.

[15] H. Al Mandhari, W. Shalish, E. Dempsey, M. Keszler, and P. Davis, "Po-0726 international survey on peri-extubation practices in extremely premature infants," 2014.

[16] A. E. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[17] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[18] M. Mueller, J. S. Almeida, R. Stanislaus, and C. L. Wagner, "Can machine learning methods predict extubation outcome in premature infants as well as clinicians?," *Journal of neonatal biology*, vol. 2, 2013.

[19] H.-J. Kuo, H.-W. Chiu, C.-N. Lee, T.-T. Chen, C.-C. Chang, and M.-Y. Bien, "Improvement in the prediction of ventilator weaning outcomes by an artificial neural network in a medical icu," *Respiratory care*, vol. 60, no. 11, pp. 1560–1569, 2015.

[20] Y. Gao, A. Xu, P. J.-H. Hu, and T.-H. Cheng, "Incorporating association rule networks in feature category-weighted naive bayes model to support weaning decision making," *Decision Support Systems*, vol. 96, pp. 27–38, 2017.

[21] N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," *arXiv preprint arXiv:1704.06300*, 2017.

[22] A. Gottschalk, M. C. Hyzer, and R. T. Geer, "A comparison of human and machine-based predictions of successful weaning from mechanical ventilation," *Medical Decision Making*, vol. 20, no. 2, pp. 160–169, 2000.

[23] A. Mikhno and C. M. Ennett, "Prediction of extubation failure for neonates with respiratory distress syndrome using the mimic-ii clinical database," in *2012 Annual international conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5094–5097, IEEE, 2012.

[24] B. Saugel, P. Rakette, A. Hapfelmeier, C. Schultheiss, V. Phillip, P. Thies, M. Treiber, H. Einwächter, A. von Werder, R. Pfab, *et al.*, "Prediction of extubation failure in medical intensive care unit patients," *Journal of critical care*, vol. 27, no. 6, pp. 571–577, 2012.

[25] J. A. Chaparro and B. F. Giraldo, "Power index of the inspiratory flow signal as a predictor of weaning in intensive care units," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 78–81, IEEE, 2014.

[26] L. J. Kanbar, C. C. Onu, W. Shalish, K. A. Brown, G. M. Sant'Anna, D. Precup, and R. E. Kearney, "Undersampling and bagging of decision trees in the analysis of cardiorespiratory behavior for the prediction of extubation readiness in extremely preterm infants," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4940–4944, IEEE, 2018.

[27] J. A. Chaparro, B. F. Giraldo, P. Caminal, and S. Benito, "Performance of respiratory pattern parameters in classifiers for predict weaning process," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4349–4352, IEEE, 2012.

[28] T. Chen, J. Xu, H. Ying, X. Chen, R. Feng, X. Fang, H. Gao, and J. Wu, "Prediction of extubation failure for intensive care unit patients using light gradient boosting machine," *IEEE Access*, vol. 7, pp. 150960–150968, 2019.

[29] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International Conference on Machine Learning*, pp. 3145–3153, PMLR, 2017.

[30] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.

[31] G. D. Perkins, D. Mistry, S. Gates, F. Gao, C. Snelson, N. Hart, L. Camporota, J. Varley, C. Carle, E. Paramasivam, *et al.*, "Effect of protocolized weaning with early extubation to noninvasive ventilation vs invasive weaning on time to liberation from mechanical ventilation among patients with respiratory failure: the breathe randomized clinical trial," *Jama*, vol. 320, no. 18, pp. 1881–1888, 2018.

[32] J. Raiten, R. H. Thiele, and E. C. Nemergut, "13 - anesthesia and intensive care management of patients with brain tumors," in *Brain Tumors (Third Edition)* (A. H. Kaye and E. R. Laws, eds.), pp. 249 – 281, Edinburgh: W.B. Saunders, third edition ed., 2012.

[33] A. Esteban, F. Frutos, M. J. Tobin, I. Alía, J. F. Solsona, V. Valverdu, R. Fernández, M. A. de la Cal, S. Benito, R. Tomás, *et al.*, "A comparison of four methods of weaning patients from mechanical ventilation," *New England Journal of Medicine*, vol. 332, no. 6, pp. 345–350, 1995.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

[35] C. Kaul, N. Pears, and S. Manandhar, "Sawnet: A spatially aware deep neural network for 3d point cloud processing," *arXiv preprint arXiv:1905.07650*, 2019.

[36] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10101–10111, 2019.

[37] C. Kaul, N. Pears, and S. Manandhar, "Fatnet: A feature-attentive network for 3d point cloud processing," *arXiv preprint arXiv:2104.03427*, 2021.

[38] C. Herzog, "Technological opacity of machine learning in healthcare," in *Weizenbaum Conference*, p. 9, DEU, 2019.

[39] D. S. Watson, J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes, and L. Floridi, "Clinical applications of machine learning algorithms: beyond the black box," *Bmj*, vol. 364, 2019.

[40] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–9, 2020.

[41] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, "Explainable machine learning in deployment," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 2020.

[42] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE, 2018.

[43] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765–4774, 2017.

[44] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104*, 2017.

[45] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[46] Y. Jia, J. Burden, T. Lawton, and I. Habli, "Safe reinforcement learning for sepsis treatment," in *8th IEEE International Conference on Healthcare Informatics*, York, 2020.

[47] I. Habli, T. Lawton, and Z. Porter, "Artificial intelligence in health care: accountability and safety," *Bulletin of the World Health Organization*, vol. 98, no. 4, p. 251, 2020.

[48] US Food and Drug Administration, "Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SAMD)—discussion paper and request for feedback. 2019," 2019.

[49] Care Quality Commission and Medical and Healthcare products Regulatory Agency, "Using machine learning in diagnostic services: A report with recommendations from CQC's regulatory sandbox," 2020.

[50] P. Theerawit, D. Eksombatchai, Y. Sutherasan, T. Suwatanapongched, C. Kiatboonsri, and S. Kiatboonsri, "Diaphragmatic parameters by ultrasonography for predicting weaning outcomes," *BMC pulmonary medicine*, vol. 18, no. 1, pp. 1–11, 2018.