# AutoWean: Extubation Failure Risk Estimation for Critically Ill Patients

Jean Park
hlpark@seas.upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

Amanda Watson
aawatson@seas.upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

Xiayan Ji
xjiae@seas.upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

Kyle C. Quinn
Kyle.Quinn@atlanticare.org
AtlantiCare Regional Medical Centers
Atlantic City, NJ, United States

James Weimer
james.weimer@vanderbilt.edu
Vanderbilt University
Nashville, TN, United States

Insup Lee
lee@seas.upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

## ABSTRACT

Mechanical ventilation is a life-saving intervention that provides breathing support for patients who cannot breathe independently, it is especially common in patients admitted to intensive care units (ICU). Extubation is the process of removing the hardware from the airway used to provide mechanical ventilation when it is no longer required. However, extubation has many potential complications, with an overall failure rate of 2-25% of patients. We present AutoWean, a system that improves the prediction of extubation outcomes in ICU patients by displaying risk levels via a feedback system provided to clinicians. Our system uses an ensemble method that combines the output of labeling functions leveraging domain knowledge from clinicians to distinguish high and low-risk patients for each risk factor. We evaluated our AutoWean model on a dataset collected over two years containing 827 extubations over 494 patients that were weaned at ICU's at the Hospital of University of Pennsylvania. The results show that patient risk can be stratified among five risk categories. The three highest risk bins indicate an extubation failure rate of over 60%, while approximately 35% for the two lowest risk bins. Most importantly, AutoWean provides decision support to clinicians attempting to delineate which borderline patients should be given a trial of extubation.

## CCS CONCEPTS

• **Applied computing → Health care information systems**.

## KEYWORDS

Medical Dataset, Machine Learning, Extubation Risk Prediction

**ACM Reference Format:**
Jean Park, Amanda Watson, Xiayan Ji, Kyle C. Quinn, James Weimer, and Insup Lee. 2022. AutoWean: Extubation Failure Risk Estimation for Critically Ill Patients. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE' 22), November 17–19, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3551455.3559601

*CHASE' 22, November 17–19, 2022, Washington, DC, USA*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9476-5/22/11…$15.00
https://doi.org/10.1145/3551455.3559601

## 1 INTRODUCTION

Breathing is a natural process that supports critical functions in the body, such as oxygenation and the removal of waste products. Mechanical ventilation is a life-saving intervention that provides breathing support for patients who cannot breathe independently. It is common to use mechanical ventilators for intensive care unit (ICU) patients, with approximately half of all ICU patients requiring mechanical ventilation [23, 4]. Intubation is the process of placing the endotracheal tube into the airway to facilitate respiratory support. Extubation is the process of removing the endotracheal tube from the airway when respiratory support is no longer required. Similarly, weaning is the gradual process of withdrawing mechanical ventilation support [1].

Extubation is a perilous process rife with potential complications. Extubation failure occurs in 2 to 25% of patients [28]. These patients experience complications such as oxygen desaturation, hypercapnea [27], fatigue, or difficult reintubations that can even lead to death. The risk of these complications increases with the duration of ventilator dependence, with a particularly marked increase in complications at greater than one week. Conversely, extubating prematurely can also cause significant complications. Thus, medically maximizing the patient prior to extubation is critical, with factors such as timing, sedation, and ventilator management being crucial to optimal patient outcomes.

Numerous medical research and clinical trials have been done to recognize critical factors for weaning a patient [33, 29, 9, 13, 30]. In general, extubation occurs after a weaning readiness test involving spontaneous breathing trials (SBT) or low levels of ventilation assistance [33]. A successful SBT indicates tolerance of unassisted breathing imposed in postextubation [33], a sign of readiness to extubate. In addition to SBT, other criteria are used to assess extubation timing. Seymour et al. [29] claim that minute ventilation recovery time post-SBT can predict extubation outcomes in ICU patients. Cohen et al. [9] found out that frequency/tidal volume ratio (f/VT) with Automatic tube compensation (ATC) at the start of the breathing trial is effective in predicting successful weaning. In additional studies [13], [30], more risk factors for extubation failures are identified.

This study aims to develop an extubation risk estimation model, AutoWean, using a majority vote to improve the prediction of extubation outcomes in ICU. The evaluation is performed on the dataset collected from the ventilators across ICUs in the Hospital of University of Pennsylvania from 2021 to 2022. Our AutoWean model provides risk levels of given extubations based on the labeling functions formulated by clinicians and medical research. Each labeling function calculates the prediction score for individual risk factors labeled high and low-risk, respectively. The calculated weights are applied to the model, outputting the risk level into five categories: high, high/med, med, med/low, and low.

The goal of our approach is to support extubation timing decisions and reduce the significant complications associated with extubation failure, i.e., reintubation within 48 hours of extubation. Our AutoWean system evaluates patients' risks through the ensembler, which combines the output of labeling functions where domain knowledge for distinguishing between high and low risk are encoded. Then, feedback is provided to the clinicians for interpretation of the risk level and extubation decision.

We perform our AutoWean model's evaluation on the dataset collected from the ventilators across ICUs in the Hospital of University of Pennsylvania from 2021 to 2022. In total, 827 extubations were extracted over 494 ICU admissions with ventilator support over those two years. We collect data from medical devices giving us access to time-series respiratory information and categorical variables. Due to the limitations of our data aquisition system, we do not have access to demographic info or ground truth labels. Regardless of this, we are able to extrapolate enough information to create a useful system. Our results show that extubations placed in three high-risk bins had an extubation failure percentage over 60%, and approximately 35% for two low-risk bins.

More specifically, our contributions are as follows:

(1) We develop the AutoWean model to improve extubation risk estimation for patients in ICU using labeling functions that leverage the domain knowledge of clinicians.
(2) We present a clinical decision support system that can aid clinician with the potential to aid clinicians in extubation timing decisions.
(3) Evaluate AutoWean system on the dataset collected from ventilators in ICUs across the Hospital of University of Pennsylvania and present result of extubation failure rate in five risk categories.

The remainder of this paper is structured as follows: Section 2 summarizes the work related to our research. In Section 3, we motivate and formulate our problem as well as describe our dataset. Section 4 details the data processing and feature extraction. Section 5 describes the the  System. Section 6 discusses the evaluation of our system and its comparisons to other state-of-the-art methods. Section 7 discusses possible future work that can be done to improve the model and labeling functions. Finally, we wrap up our paper with a conclusion in Section 8.

## 2 RELATED WORK

Weaning a patient from a ventilator is one of the most common, important, and at time frustrating challenges for the clinician in the ICU. Recently, researchers have been attempting to decrease the extubation failure rate in ICU by analyzing extubated patient data and applying machine learning to predict extubation outcomes. This section discusses the machine learning and clinical decision support systems in healthcare and examples of such systems used for extubation prediction.

### 2.1 Machine Learning in Healthcare

Today, technological evolution in medicine has led to the production and collection of massive amounts of healthcare data[19]. This data is utilized for better treatment [3], personalized medicine [8], etc [26]. With analysis and interpretation of this big data, machine learning algorithms can distinguish the pattern in patient data and predict the outcome or treatment based on their individual attributes [10], [25]. In certain applications [11], ML-based prediction models, or AI, outperform human's ability to process, analyze, and find patterns in complex, high-dimensional data [16]. In turn, machine learning techniques are adopted in diverse applications in healthcare, and the number of such applications is increasing every year [18],[2].

Clinical decision support systems (CDSS) are intended to provide clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance patient treatment and outcomes [25],[32],[34]. These systems widely adopt machine learning tools as a method of providing individualized recommendations. CDSS directly aid in clinical decision-making, in which the characteristics of an individual patient are matched to a computerized clinical knowledge base, and patient-specific assessments or recommendations are then presented to the clinicians or the patient for a decision [31]. CDSS leverage data that is already collected and stored in electronic health records (EHR) and research databases [7]. These systems are made to support the clinician. Thus, in real clinical scenarios, information must be easy to access and interpret. Then, clinicians can fuse the CDSS recommendations with any additional contextual information that they acquire from direct interaction from the patient [20]. In its best instantiation, CDSS is synergistically combined with clinicians' insight and high-level cognitions to create high quality medical care. In this work, we strive to create a CDSS that aids clinician decisions in patients that are difficult to extubate.

### 2.2 Extubation Outcome Prediction

Fabregat et al. [12] used classification learners on the patient data from ICU in Spain given a dataset of 697 extubations. In total this dataset only had 50 extubation failures. Their dataset contains 20 predictors including time series variables from monitoring devices, demographics, medical records, and respiratory logs. They tested Logistic Discriminant Analysis (LDA), Gradient Boosting Method (GBM), and Support Vector Machines (SVM) and found that the SVM was the most accurate classifier with an accuracy of 94.6%. Overall this method shows high accuracy over their data. However, this work is limited by their small dataset. With the low number of extubations, they used 20 minutes binning over 2 hour period before extubation. Each bin become a new datapoint creating 4182 data points or six times the original number of extubations. While we were not able to get access to their dataset or model to verify, it is possible that extubations from the same patient were placed into both the training and test data. This could create bias in their model and analysis so future work will be needed to compare to their method.

Chen et al. [6] developed a Light Gradient Boosting Machine (LightGBM) model for extubation failure prediction on 3636 patients from the freely accessible clinical data MIMIC-III [17]. Initially, 68 features were evaluated based on their average prediction result with the top 36 features being chosen for the final model. These features include patient demographics, vital signs, laboratory measurements, ventilator information, clinical intervention, and clinical scores. Similar to our research, they do not have direct intubation and extubation information, so they inferred patients' extubation with ventilation parameters associated with the usage of ventilators indicated a status change from intubated to not intubated. Overall, their model showed a high accuracy of 80.2%. But their analysis was limited as they do not use bins for the time series data and take values outside of the intubation period. In addition, they imputed the missing value of the features with mean or median value. Only nine of the 68 features had 0% missing degrees, whereas all other features' missing degrees were 0-40%. On the other hand, features used for our extubation risk estimation model did not have any missing values.

# 3 PRELIMINARIES AND PROBLEM FORMULATION

In this section, we describe the dataset used for training and evaluation in this paper, the labeling functions which encode clinical insights, and the problem formulation. First, we will describe our dataset. Then we will discuss the clinical knowledge encoded as labeling functions. Finally, we will present the problem.

## 3.1 Dataset

Of a study approved by the Institutional Review Board of the University of Pennsylvania (IRB #851405). In this study, medical device data was collected from our data acquisition system, VitalCore [7], a medical device integration platform. VitalCore is connected to 247 ventilators from four different vendors in ICUs across the Hospital of University of Pennsylvania. These ventilators send data in an HL7 message format [15] to the servers in VitalCore every minute, where it is stored for future analysis. Overall, we were able to collect 827 extubations belonging to 494 patients over the course of two years. Among the 827 extubations, 373 of them are extubation failures while 454 are extubation successes. Overall, we have a higher extubation failure rate of 45% compared to an average of 10-20% in ICU [24]. We attribute the difference to the timeline of our collected dataset from 2021 to 2022, during the COVID-19 pandemic.

As the treatment of COVID-19 evolved so did its protocols. At times this led to frequent extubation attempts for patients on ventilators. We identified patients who were extubated significantly more than the average number. In particular, eight patients were extubated more than ten times accounting for 2-5 which occurs over half of the patients with multiple extubation attempts [21]. In total, 90 out of 99 extubations were failures, increasing the extubation failure rate significantly by approximately 6%. Regardless, we include them as valid extubations in our dataset.

Before building and testing our model, we set aside a hold out set. Thus, we split our dataset $Z = (X, Y)$ into two portions $Z_1 = (X_1, Y_1)$ and $Z_2 = (X_2, Y_2)$. $Z_1$ accounts for approximately 87% of our total dataset and is used for training and testing the models. $Z_2$ accounts for approximately 13% of our data and will be used during evaluation as a hold-out dataset. The percentage of the dataset being used as a hold-out set is not perfectly 10% as we ensure patients with multiple extubations only appear in one of $Z_1$ or $Z_2$. Across the two portions of our datasets, we observe a similar rate of extubation failure at 44% in $Z_1$ and 41% in $Z_2$.

## 3.2 Labeling Functions

We collected labeling functions from clinician knowledge, published medical research, and basic data analysis. These labeling functions should either distinguish a patient as high risk or low risk. For example, a patient being on a ventilator for over 72 hours is an example of a high risk labeling function and total ventilation time less than 24 hours is an example of a low risk labeling function. If a patient is not classified as high or low risk for any specific feature, they can be thought of as unknown risk. This can occur when data is missing or when data is outside of the scope of the labeling functions. Overall, we collected sixteen labeling functions from clinicians, published research, and basic analysis techniques.

## 3.3 Problem Formulation

Our input space $\mathcal{X}$ encompasses the dataset described above such that every $x \in \mathcal{X}$ describes a feature collected from our medical devices. In addition, we have a label space $\mathcal{Y} = \{-1, 1\}$ where 1 represents an extubation failure and -1 represents a successful extubation. Clinically, extubation is considered a failure if the patient is reintubated within 48 hours of the extubation. Thus, our ground truth labeling function can be described as $f_{gt} : \mathcal{X} \mapsto \mathcal{Y}$.

Next, we collected and evaluated a set of labeling functions from clinicians, expert research, and data analysis techniques. These labeling function distinguish high and low risk factors for extubation failures. They use a label space $\bar{\mathcal{Y}} = \{high, low\}$ corresponding to two conclusive and mutually exclusive risk levels of extubation failure: high and low. To make the system intuitive, we extend our label space to $\hat{\mathcal{Y}} = \{high, low, unknown\}$ by adding a label for unknown risk. Thus, we have two sets of binary partial label functions: (1) a set $H$ such that $\forall f_h \in H$, $f_h : \mathcal{X} \mapsto \hat{\mathcal{Y}}_h$ where $\hat{\mathcal{Y}}_h = \{high, \neg high\}$ such that $\neg high = \{low, unknown\}$ and (2) a set $L$ such that $\forall f_l \in L$, $f_l : \mathcal{X} \mapsto \hat{\mathcal{Y}}_l$ where $\hat{\mathcal{Y}}_l = \{low, \neg low\}$ such that $\neg low = \{high, unknown\}$.

The goal of this paper is to evaluate each patients' risk for failing extubation. By using $X$, $Y$, $H$ and $L$, we seek a labeling function $f^* : \mathcal{X} \mapsto \hat{\mathcal{Y}}$ such that $f^*$ the probability of a extubation failure occurring decreases from the high to low-risk category. Additionally, we seek to create a risk stratification that provides decision support to clinicians on the patients that are borderline. In the following section, we introduce AutoWean, a solution to the problem formulation we just described.

# 4 DATA PROCESSING AND FEATURE EXTRACTION

In this section, we describe the inclusion criteria and algorithms we used to process and select our data. Then, we describe the feature extraction process.

## 4.1 Data Processing

We process our data in four steps. First, we detect if a patient is connected to the ventilator to identify potential intubation and extubation times. Second, we define and determine periods of reduced pressure (PRP's) that estimate clinical SBT's. Third, we label ground truth as extubation pass or failure. Fourth, we discuss and remove special cases from the dataset.

**Patient Detection** Ventilators may still send HL7 messages when they are not connected to patients. However, as disconnected ventilators do not contain tidal volume in their HL7 messages, we use it to distinguish ventilator states. The absence of tidal volume means that patient is not connected to the ventilator. Moreover, a ventilator may be connected to a different patient when a patient is weaned and never intubated again. While a ventilator does not send any patient identification information, a monitoring device located in the same room sends HL7 messages with a patient identification number to the VitalCore system. We use this to distinguish patients. Given the extracted list of tidal volume belonging to one patient, we compute potential intubation and extubation time periods. In the next step we use these to determine signs of a PRP.

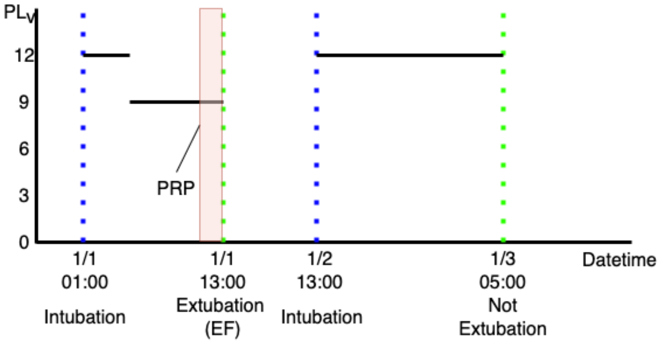


**Figure 1: Example of Identifying Periods of Reduced Pressure and Distinguishing Planned Extubations**

**Algorithm 1** Compute PRP

```
Input: Pressure Level List $PL_v$, Ventilator Pressure Level Timestamp List $PL_t$, List of Potential Intubation and Extubation Time $D$
Output: List of PRP Start Time, PRP End Time, Intubation Start Time, Extubation End Time $P$

Initialize empty list $P$
Initialize variables $l_{idx} \leftarrow 0$
for $\{s, e\}$ in $D$ do
    Initialize empty list $L$
    while $l_{idx} < length(PL_t)$ & $PL_t[l_{idx}] < s$ do
        $l_{idx} \leftarrow l_{idx} + 1$
    end while
    while $l_{idx} < length(PL_t)$ & $PL_t[l_{idx}] \geq s$ & $PL_t[l_{idx}] < e$ do
        $L \leftarrow L \bigcup \{l_{idx}\}$
        $l_{idx} \leftarrow l_{idx} + 1$
    end while
    if $7 \leq PL_v[l_{idx} - 1] \leq 10$ then
        $k \leftarrow l_{idx} - 1$
        while $PL_v[l_{idx} - 1] = PL_v[k]$ & $k > L[0]$ do
            $k \leftarrow k - 1$
        end while
        Set PRP start time $prp_s \leftarrow PL_t[k + 1]$
        Set PRP end time $prp_e \leftarrow PL_t[l_{idx} - 1]$
        if $prp_e - prp_s \geq 2$ hours then
            $l \leftarrow l_{idx} - 1$
            while $prp_e - PL_t[l] < 2$ hours do
                $l \leftarrow l - 1$
            end while
            Update PRP time $prp_s \leftarrow PL_t[l + 1]$
        end if
        if $prp_e - prp_s > 30$ minutes then
            $P \leftarrow P \bigcup \{prp_s, prp_e, s, e\}$
        end if
    end if
end for
return $P$
```

**Periods of Reduced Pressure** The Hospital of University of Pennsylvania uses a guideline to evaluate patients for potential extubation. This guideline ensures clinicians take into account SBT's and a number of other factors. During an SBT, clinicians decrease the amount of pressure used to support patients' breathing. This is done to assess the patients' ability to generate sufficient pressure to breathe on their own if extubated. Normally, medical professionals record SBT initiation and termination time in EHR (Electronic Health Records), but we do not have access to this data through our system, only medical device data. Accordingly, we estimate SBT's with pressure level change in ventilator settings. Before each extubation, we validate if the pressure level decreased from a higher level to anything from seven to ten and remained steady for at least 30 minutes to estimate the SBT. We refer to this phase as a period of reduced pressure (PRP). Only when there is a PRP prior to potential extubation do we consider them as real extubation. Further details for finding PRPs of an individual patient is shown in Algorithm 1 and Fig. 1.

**Ground Truth Label** Ventilator disconnection for greater than ten continuous minutes likely indicates one of the following:

(1) Extubation
(2) Transference to a travel ventilator for offsite imaging or studies
(3) Undergoing of invasive procedures that stop mechanical ventilation without extubation

To distinguish planned extubation from unplanned or non-extubations, we take periods where tidal volume was absent for more than ten minutes and trace back to check if PRP was done immediately prior to ventilation disconnection. That is, we only consider extubations in which clinicians evaluate patients' breathing ability. Once we obtain actual extubations from Algorithm 1, the ground truth becomes 1 when the time that takes to reintubate is less than 48 hours; otherwise -1.

We label extubation result as a failure if the same patient is reintubated within 48 hours of previous extubation, otherwise success. However, in some circumstances, it could be hard to determine whether extubation was successful or not. For instance, patients intubated for more than two weeks are presumed to have tracheostomy tubes. Accordingly, we remove all of their extubations from our dataset as tracheostomy is often considered for prolonged mechanical ventilation and failed extubation. Furthermore, some patients were extubated in different facilities. Given our limited data, it is hard to discern their actual extubation result in previous facilities as they could have been on a transport ventilator to move their location. Thus, we only take into account extubations in the last facility.

## 4.2 Feature Extraction

In our dataset, we compute categorical variables such as a previous number of extubations and total ventilation duration time, and time-series variables like respiratory rate (RR) abnormal time and rapid shallow breathing index (RSBI). Using Algorithm 1 we can extract every extubation belonging to each patient and calculate the previous number of extubations that occurred prior to it. Also, we can easily compute the total ventilation time. No further processing is required for categorical data and we can directly use the value for our model.

On the contrary, time-series variables have abundant and redundant data points as they are sent every minute during patient intubation. To capture information most relevant to extubation outcomes, data points within PRP have prognostic values of weaning results [35]. It is considered that patients are likely to fail extubation if the patient's respiratory rate during SBT is above 35 breaths per minute for more than five minutes [5]. Therefore, we examine events where the respiratory rate is abnormal, i.e., over 35 bpm, and record its total time. Furthermore, we compute RSBI by taking the average respiratory rate(RR) and tidal volume(VT), using Equation 1.

$$RSBI(bpm/L) = \frac{RR(bpm)}{VT(L)} \tag{1}$$

# 5 AUTOWEAN SYSTEM

The Autowean system is a clinical decision support system that evaluates and provides feedback on a patient's risk of extubation failure. It leverages domain knowledge from clinicians followed by machine learning techniques to identify patients who may not tolerate extubation. Once the risk level has been determined, it provides feedback to the clinician who will make decisions regarding patient care. It is comprised of three components: feature evaluator, ensembler, and feedback system. The feature evaluator evaluates each labeling function to determine its usefulness to the ensembler. The ensembler uses machine learning techniques to combine the outputs of the labeling functions. The feedback system displays the risk level to the clinician for interpretation and clinical decision support. An overview of the AutoWean system is depicted in Figure 2.

## 5.1 Labeling Function Evaluator

Many of our labeling functions are designed by clinicians or have been evaluated through medical research. Furthermore we use data analysis techniques to create additional labeling functions. Thus, before we use a labeling function in our system we evaluate it to determine if it is a good predictor of extubation success or failure. To do this we use $Z_1 = (X_1, Y_1)$ which accounts for approximately 87% of our total dataset. In total that gives us 317 extubation failures and 406 extubation successes to evaluate our labeling functions. With each tested labeling function, we calculate a prediction score as follows:

Given a labeling function $f \in F$, we calculate its prediction score over all $x \in X_1$ as

$$s_f = \frac{\sum_{x \in X_1} \mathbf{1}(f(x) = 1 \wedge f_{gt}(x) = 1)}{\sum_{x \in X_1} \mathbf{1}(f(x) = 1)} \tag{2}$$

i.e. the percentage of patients that fail extubation out of the patients labeled by the labeling function. If a labeling function has a very high score it is a strong predictor of extubation failure. We will denote this set of lableing functions as $H$. If a labeling function has a very low score it is a strong predictor of extubation success. We will denote this set of labeling functions as $L$. If a labeling function has a score in the middle, it is not a strong predictor of success or failure. We remove these labeling functions before the next step.

## 5.2 AutoWean Model

The AutoWean model leverages an ensemble approach to combine the outputs of both high and low risk labeling functions. Specifically it uses a weighted majority to classify the risk level for each patient. It is trained using data from $Z_1$ and will be evaluated in a later section via cross validation as well as using $Z_2$ as a hold out set. We show the overall algorithm used to generate this model in Algorithm 2.

Given a labeling function $f_h \in H$, its weight $w_{f_h}$ is

$$w_{f_h} = \frac{\sum_{x \in X_1} \mathbf{1}(f_h(x) = high \wedge f_{gt}(x) = 1)}{\sum_{x \in X_1} \mathbf{1}(f_h(x) = high)} \tag{3}$$

i.e. the percentage of patients that actually go on to fail extubation out of the patients labeled as high risk by $f_h$.

Similarly, given a labeling function $f_l \in L$, its weight $w_{f_l}$ is

$$w_{f_l} = \frac{\sum_{x \in X_1} \mathbf{1}(f_l(x) = low \wedge f_{gt}(x) = 1)}{\sum_{x \in X_1} \mathbf{1}(f_l(x) = low)} \tag{4}$$

i.e. the percentage of patients that go on to fail extubation out of the patients labeled as low risk by $f_l$. We can see that the weights $w_{f_h}, w_{f_l} \in [0, 1]$ as they are percentages. Moreover, the weights measure a labeling functions actual risk using the ground truth $f_{gt}$ in the dataset $X_1$. For every $f_h \in H$, as $w_{f_h}$ approaches 1, the calculated risk level increases. On the other hand, for every $f_l \in L$, as $w_{f_l}$ approaches 0, the calculated risk level decreases.

Next, we compute the majority vote by applying the following equation to each $x \in X$:

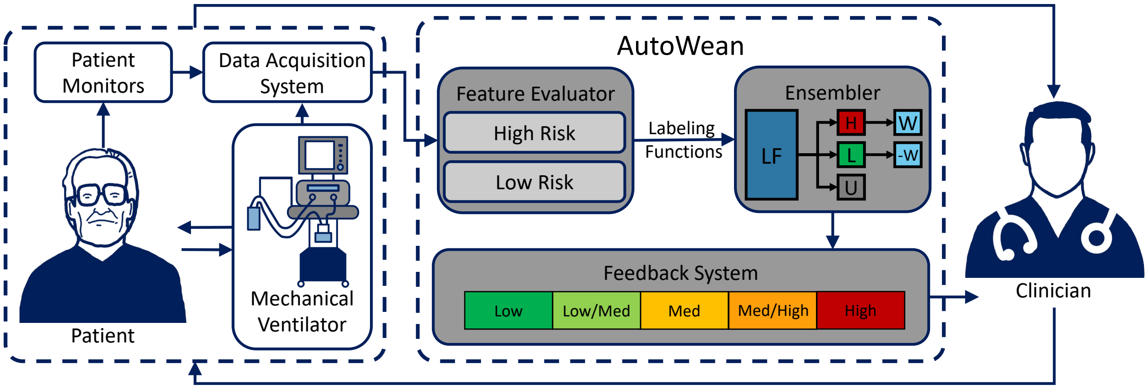


**Figure 2: System Architecture**

**Algorithm 2** AutoWean Model Algorithm

```
Input: Training Dataset $Z_1 = (X_1, Y_1)$, High-Risk Labeling Functions $H$, Low-Risk Labeling Functions $L$, Thresholds $t_l$, $t_{lm}$, $t_m$, $t_{mh}$
Output: Risk Estimation Function $f^*$
Obtain the ground truth labeling function $f_{gt}$ from $Z_1$
Initialize empty lists $w_L$, $w_H$, $s$
for $f_h$ in $H$ do
    $w_{f_h} = \frac{\sum_{x \in X_1} 1(f_h(x)=high \wedge f_{gt}(x)=1)}{\sum_{x \in X_1} 1(f_h(x)=high)}$
    $w_H.append(w_{fh})$
end for
for $f_l$ in $L$ do
    $w_{f_l} = \frac{\sum_{x \in X_1} 1(f_l(x)=low \wedge f_{gt}(x)=1)}{\sum_{x \in X_1} 1(f_l(x)=low)}$
    $w_L.append(w_{fl})$
end for
for $x$ in $X_1$ do
    $s_x = c_h \sum_{w_{fh} \in w_H} w_{f_h} \mathbf{1}(f_h(x) = high) - c_l \sum_{w_{fl} \in w_L} w_{f_l} \mathbf{1}(f_l(x) = low)$
    s.append($s_x$)
end for
for $s_x$ in $s$ do
    Define $f^*(x)$ as $\begin{cases} low & s_x < t_l \\ low/med & t_l \le s_x < t_{lm} \\ med & t_{lm} \le s_x < t_m \\ med/high & t_m \le s_x < t_{mh} \\ high & s_x \ge t_{mh} \end{cases}$
end for
return $f^*$
```

$$s(x) = c_h \sum_{f_h \in H} w_{f_h} \mathbf{1}(f_h(x) = high) - c_l \sum_{f_l \in L} w_{f_l} \mathbf{1}(f_l(x) = low) \tag{5}$$

The sum $s(x) \in R$ shows an estimated numerical value of extubation failure risk: the larger $s(x)$, the higher risk the patient has.

Finally, we will pick two thresholds $t_{lm}$ and $t_{mh} \in R$ to split the low, medium and high risk levels based on the $s(x)$ computed. The final extubation failure risk estimation function $f^* : \mathcal{X} \mapsto \hat{\mathcal{Y}}$ is

$$f^*(x) = \begin{cases} low & s_x < t_l \\ low/med & t_l \le s_x < t_{lm} \\ med & t_{lm} \le s_x < t_m \\ med/high & t_m \le s_x < t_{mh} \\ high & s_x \ge t_{mh} \end{cases} \tag{6}$$

## 5.3 Feedback System

Once the risk level has been determined, we convey this information to the clinicians. This rating system simplifies the complex data from a large number of source to an easily understood and quickly conveyed assessment of risk. Clinically there are patients that need no further evaluation and on a cursory exam, one is able to decide whether they are extubatable or not. The clinical utility of this system comes from the stratification of borderline cases. This is shown by the low/med, med, and med/high bins. It provides a validated decision support structure that is easy to reference when justifying clinical decisions, particularly for these challenging patients.

| | Extubation success | Extubation failure |
|---|---|---|
| Low risk | TN | FN |
| High risk | FP | TP |

**Table 1: The four possible outcomes.**

## 5.4 Theoretical Guarantee

In this section, we illustrate the theoretical guarantee of our system. It provides the upper bound on two important error metrics, namely, false negative rate (FNR) and false positive rate (FPR) for clinical decisions. This provides clinicians with confidence in the accuracy of the predictions of the system. Specifically, we provide a Probably Approximately Correct (PAC) [14] guarantee on the probability of the clinical decision being correct, which is achieved by stratifying the sum $s(x)$ with two PAC thresholds $\tau_{\text{fp}}$ and $\tau_{\text{fn}}$ as in [22]. In Table 1, we listed out the four possible results for extubation based on the risk prediction:

Accordingly, False Negative Rate (FNR) and False Positive Rate (FPR) are defined as below:

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \qquad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

In other words, FN means we predict a patient that has a high likelihood of failing extubation to have low risk, and FP means we consider a patient that will succeed in the extubation to have high risk. Both FN and FP are undesirable, and we should minimize or control them at a low level for our system to be helpful.

Therefore, we provide a PAC that both the FNR and FPR are lower than a predefined level $\epsilon$ with a probability of $1 - \delta$ [22]. In particular, we first split the $s(x)$ in the training dataset $Z_1$ into two subsets based on extubation success and failure. On the failure subsets, we sort the $s(x)$ in ascending order, and choose the score located at the $k_{\text{fn}}$-th position as the threshold $\tau_{\text{fn}}$. The position choice enforces the $\epsilon$ constraint for FNR with probability $1 - \delta$. Similarly, we use the score at the $k_{\text{fp}}$-th position on the success subset as $\tau_{\text{fp}}$, which guarantee FPR with probability $1 - \delta$. Since $\tau_{\text{fn}}$ is calucalted from the failure subsets with higher $s(x)$ than that of the success subsets, it is reasonable to assume $\tau_{\text{fn}} > \tau_{\text{fp}}$. Then, the two thresholds guide our decision as follows:

$$f^*(x) = \begin{cases} low, & s(x) < \tau_{\text{fp}} \\ med, & \tau_{\text{fp}} \leq s(x) \leq \tau_{\text{fn}} \\ high, & s(x) > \tau_{\text{fn}} \end{cases} \tag{7}$$

Specifically, if we let $t_{mh} = \tau_{\text{fn}}$ and $t_{lm} = \tau_{\text{fp}}$, we guarantee that:

$$P[\text{FNR} \leq \epsilon] \geq 1 - \delta$$
$$P[\text{FPR} \leq \epsilon] \geq 1 - \delta$$

In summary, using the two thresholds $t_{mh}$ and $t_{lm}$ on the sum $s(x)$, we theoretically guarantee the accuracy of the clinical decision is higher than $1 - \epsilon$ with a confidence of $1 - \delta$. To validate the effectiveness of our theoretical guarantee, we empirically evaluate its performance in the following Evaluation section.

## 6 EVALUATION

In this section, we evaluate the performance of our extubation risk estimation model. We present our labeling functions predictive of the risk and their weights for each risk level. Next, we show three different hyperparameters used with our labeling functions and compare their result. Then, we display the cross-validation result of our model based on the appropriate selection of the hyperparameter and the weights. Finally, we perform validation of our model over the holdout set.

For the purposes of the analysis in section, we will further split $Z_1$ into a training and testing set. $Z_3 = (X_3, Y_3)$ will be used for training and accounts for 70% of our total data. $Z_4 = (X_4, Y_4)$ is the testing dataset comprised of 17% of our dataset. $Z_2 = (X_2, Y_2)$ is used as a hold-out set and makes up 13% of our dataset. Since many patients have multiple extubations, we divide our dataset while ensuring that each patient only appears in one of $Z_1$, $Z_2$, $Z_3$, or $Z_4$. We show a phenotyping for these datasets including their folds in our cross validation (CV) in Table 2.

| Risk Factor | High risk | | Low risk | | Unlabeled | |
|---|---|---|---|---|---|---|
| | W | # | W | # | W | # |
| Previous number of extubation | 75 | 163 | 30 | 444 | 54 | 116 |
| RR abnormal time | 67 | 51 | 44 | 463 | 38 | 209 |
| RSBI | 74 | 19 | 45 | 85 | 42 | 619 |
| Total ventilation time | 67 | 225 | 29 | 355 | 45 | 143 |

**Table 3: Labeling Function. W: Weight, #: Total Number of Extubations in the Risk Level, RR: Respiratory Rate, RSBI: Rapid Shallow Breathing Index**

### 6.1 Labeling Function Accuracy

Based on domain knowledge, we present labeling functions capable of identifying the risk level for each risk factor. The list of weights and numbers of extubations for given risk factors are displayed in Table 3. Extubations in the high-risk group are likely to be failed, whereas in the low-risk are likely to be passed. The weights indicate the percentage of extubation failure out of the total number of extubations. Patients that did not meet the criteria of either high or low-risk are classified as unlabeled as shown in Table 4. For instance, 78% of patients extubated more than or equal to two times were reintubated within 48 hours. Conversely, about 70% of patients without prior extubation experienced successful liberation from the ventilator, i.e., only 30% failed extubations. Moreover, we do not label the patients who were extubated once, as the outcome is around half and therefore, unpredictable.

We designed labeling functions with an accuracy of over 65% for high-risk and less than 45% for low-risk. We only have four labeling functions as various types of ventilators across the Hospital of University of Pennsylvania had varying functionalities and did not support collecting some risk factors that clinicians deemed important. However, most of our labeling functions, except for *RSBI*, cover more than two-thirds of the extubations. Moreover, the high predictive power of individual labeling functions reduces the noise of the model during evaluation and aims to provide consistent risk stratification between different dataset.

### 6.2 Hyperparameter Tuning

Table 5 shows the list of hyperparameter configurations that determine how overall voting works in favor of high and low-risk weights. When $c_h = c_l = 1$, low-risk labeling functions are added to the total vote, as well as high-risk labeling functions. On the other hand, if the configuration is $c_h = 1$ and $c_l = -1$, the low-risk votes are subtracted from the high-risk votes. The inverse of the low-risk factors subtracted from 100 derives a similar effect as $c_h = c_l = 1$ when subtracted from overall votes. Of the three hyperparameters in Table 5, the first and third rows showed the same results. Overall, subtracting the low-risk factors performs the best out of all three settings, achieving high extubation failure estimation for the three

**Table 2: Phenotyping on dataset. EF: Extubation Failure, ES: Extubation Success, P: Patient, $I_{avg}$: Average of Total Intubation Time, ICU: Intensive Care Unit, CCU: Critical Care Unit, HVICU: Heart and Vascular ICU, MICU: Medical Intensive Care Unit, NTSICU: Neuroscience Trauma Surgical ICU, SICU: Surgical ICU, CV: Cross Validation**

| Evaluation | Dataset | EF #(%) | ES #(%) | P # | $I_{avg}$ | CCU/HVICU | ICU | MICU | NTSICU | SICU |
|---|---|---|---|---|---|---|---|---|---|---|
| CV1 | Train | 244 (43%) | 320 (57%) | 285 | 31.1 | 122 | 18 | 68 | 94 | 18 |
| CV1 | Test | 73 (46%) | 86 (54%) | 89 | 32.1 | 44 | 29 | 25 | 60 | 1 |
| CV2 | Train | 238 (42%) | 326 (58%) | 285 | 32.5 | 205 | 56 | 113 | 165 | 25 |
| CV2 | Test | 79 (50%) | 80 (50%) | 89 | 27.2 | 44 | 27 | 34 | 44 | 10 |
| CV3 | Train | 269 (46%) | 322 (54%) | 285 | 31.9 | 186 | 79 | 115 | 188 | 23 |
| CV3 | Test | 48 (36%) | 84 (64%) | 89 | 29.1 | 63 | 4 | 32 | 21 | 12 |
| CV4 | Train | 267 (45%) | 327 (55%) | 285 | 30.5 | 194 | 74 | 122 | 173 | 31 |
| CV4 | Test | 50 (39%) | 79 (61%) | 89 | 35.3 | 55 | 9 | 25 | 36 | 4 |
| CV5 | Train | 250 (43%) | 329 (57%) | 286 | 30.8 | 206 | 69 | 116 | 161 | 27 |
| CV5 | Test | 67 (47%) | 77 (53%) | 88 | 33.5 | 43 | 14 | 31 | 48 | 8 |
| Hold Out | Validation | 56 (41%) | 48 (59%) | 50 | 33.9 | 30 | 16 | 34 | 10 | 14 |

highest risk bins. In addition, it retains the lowest false-negative rate; that is, the extubation failure rate is low. Minimizing the extubation failure rate in low-risk bins is critical as it reduces the number of extubating unprepared patients and prevents serious complications.

## 6.3 5-Fold Cross Validation

We used the labeling function weights in Table 3 and the hyperparameters $c_h = 1$ and $c_l = -1$ to perform a 5-fold cross-validation on the training dataset. Among 494 patients, we randomly assigned approximately 10% of patients to the holdout set and the rest to the training set in which the number of patients has equally partitioned into five sets. As a patient can be reintubated multiple times, the number of extubations varies for each set. We use four sets for training and one for testing when evaluating each fold. When we train on all of $Z_1$ we get the risk stratisfication shown in Figure 3.

Furthermore, we provide the phenotyping across cross-validation folds and hold out set in Table 6. Hold out set $Z_2$ was not included in Table 3 and cross-validation. The ratio of extubation failure vs. success differs across the folds, but the three highest risk categories always have higher extubation failure while the two lowest risk categories have higher success. In general, the model performance meets our expectations in which the extubation failure percentage drops as the risk level decreases.

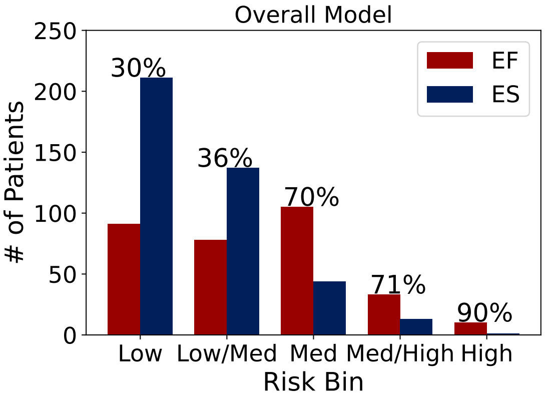


**Figure 3: Overall Model. Rate of extubation failure is given as a percentage for each bin.**

## 6.4 Hold Out Evaluation

Finally, we evaluated our model with the holdout dataset containing 50 patients with 104 extubations. Among these extubations, there were 56 failures over 17 patients. As shown in Table 6, extubations in all risk categories showed high predictability of extubation failures. It is worth noting that no patients were classified as high-risk. The failure percentage in medium-risk (62.5%) was slightly lower

| Feature | High Risk | Low Risk | Unknown |
|---|---|---|---|
| Previous number of extubation | $x \geq 2$ | $x = 0$ | $x = 1$ |
| RR abnormal time (min) | $x > 10$ | $x = 0$ | $0 < x \leq 10$ |
| RSBI | $90 \leq x \leq 200$ | $60 < x < 90$ | $(x \leq 60) \mid (x > 200)$ |
| Total ventilation time (hr) | $x > 72$ | $x < 24$ | $24 \leq x \leq 72$ |

**Table 4: Labeling Function Criteria**

| Risk Weighting | Hyperparameter | High risk | | High/Med risk | | Medium risk | | Med/Low risk | | Low risk | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | W | # | W | # | W | # | W | # | W | # |
| Labeling Function | $c_h = 1, c_l = 1$ | 66.7 | 15 | 70.5 | 78 | 36.5 | 271 | 36.4 | 154 | 47.6 | 42 |
| Labeling Function | $c_h = 1, c_l = -1$ | 85.7 | 14 | 76.7 | 86 | 61.0 | 82 | 33.3 | 198 | 27.2 | 184 |
| Labeling Function | $c_h = 1, c_l = -(\frac{100}{\sum_{f_l \in L} w_{f1} 1(f_l(x)=low)} - 1)$ | 66.7 | 15 | 70.5 | 78 | 36.5 | 271 | 36.4 | 154 | 47.6 | 42 |

**Table 5: Weighting Methods and Hyperparameter Tuning Results**

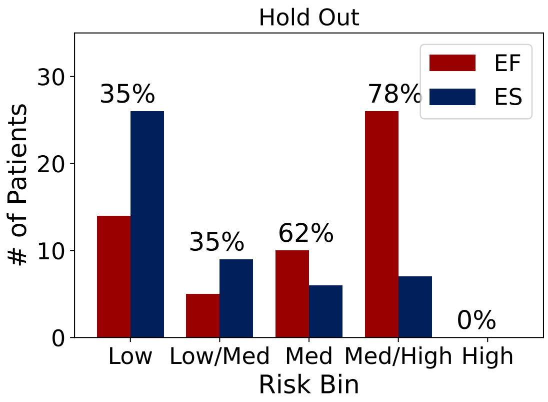


**Figure 4: Hold Out Risk Stratification. Rate of extubation failure is given as a percentage for each bin.**

than the average of 67.6% and low-risk was higher than all cross-validation results. Nonetheless, our model is capable of providing a consistent prediction of extubation outcomes for each risk-level.

## 6.5 Guarantee Validation

We first use a labeled training set for calibration, i.e., computing the two thresholds $\tau_{\text{fn}}$ and $\tau_{\text{fp}}$, and then test whether the guarantee holds on a test set. Then, we repeat the experiment for 1000 Monte Carlo trials. In each trial, we compose an i.i.d. calibration set containing 20% of the data from the labeled training set and use the rest as a test set to validate the guarantee. Finally, we evaluate the thresholds we get from the 1000 trials on a separate hold-out set $Z_2$ that is never exposed to it. We confirm the average FNR and FPR are below the average $\epsilon$ to validate the guarantee on the unseen hold-out set. Additionally, we keep track of the ratio $\hat{\delta}$ where the expected error is smaller than $\epsilon$ to validate whether the $\delta = 0.2$ constraint is met. Besides FNR and FPR, we also compute the weighted combination of the two and denote it as ERR. Compared with a threshold that maximizes the F1 score $\tau_{\text{f1}}$, our PAC thresholds not only provide a theoretical guarantee but also achieve lower empirical FNR, FPR, and ERR.

The result is shown in Table 7. For the $\epsilon = 0.37$ constraint, both FNR and FPR are smaller than 0.37, and hence we satisfy the error constraint. On the other hand, we can see that $\hat{\delta} = 0.12$ is below $\delta = 0.20$, indicating that we satisfy the confidence constraint we impose when calculating the thresholds. Notice that both thresholds lead to smaller FNR than FPR, which is desirable since FN is a more severe outcome than FP. On top of that, we have a smaller FPR than the baseline, suggesting that our method is less likely to cause alarm fatigue than the baseline.

**Table 7: Comparing our result against the baseline threshold $\tau_{\text{f1}}$, we satisfy the guarantee and have smaller error.**

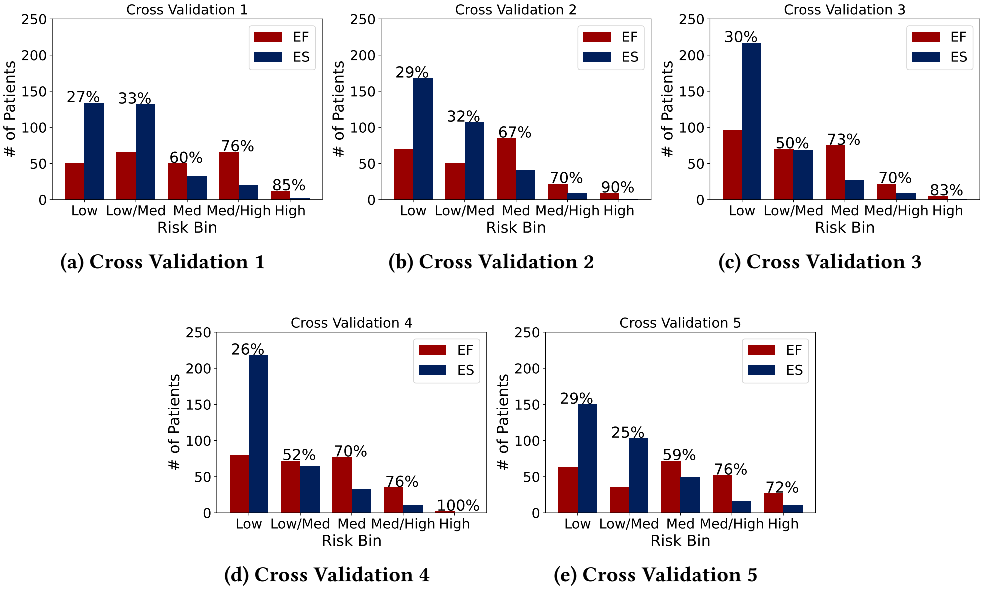
| | $\epsilon$ | $\delta$ | FPR | FNR | ERR | $\hat{\delta}$ |
|---|---|---|---|---|---|---|
| Ours | 0.37 | 0.20 | **0.32** | 0.23 | **0.25** | 0.12 |
| Baseline | - | - | 0.37 | 0.23 | 0.31 | - |

# 7 DISCUSSION AND FUTURE WORK

In this section, we discuss future work to improve our model. First, auto-generation of labeling functions can be developed to find the high and low-risk thresholds automatically. Furthermore, more risk predictors that are prognostic of extubation outcomes can be

| Dataset | Low Risk | | Med/Low Risk | | Medium Risk | | High/Med Risk | | High Risk | |
|---|---|---|---|---|---|---|---|---|---|---|
| | W | # | W | # | W | # | W | # | W | # |
| CV1 | 27.2 | 184 | 33.3 | 198 | 61.0 | 82 | 76.7 | 86 | 85.7 | 14 |
| CV2 | 29.4 | 238 | 32.3 | 158 | 67.5 | 126 | 71.0 | 31 | 90.0 | 10 |
| CV3 | 30.7 | 313 | 50.72 | 138 | 73.5 | 102 | 71.0 | 31 | 83.3 | 6 |
| CV4 | 26.8 | 298 | 52.5 | 137 | 77.0 | 110 | 76.1 | 46 | 100.0 | 2 |
| CV5 | 29.6 | 213 | 25.9 | 139 | 59.0 | 122 | 76.5 | 68 | 73.0 | 37 |
| Hold Out | 35.0 | 40 | 35.7 | 14 | 62.5 | 16 | 78.8 | 33 | - | - |

**Table 6: 5-Fold Cross Validation and Hold Out Evaluation**



**(a) Cross Validation 1** **(b) Cross Validation 2** **(c) Cross Validation 3**

**(d) Cross Validation 4** **(e) Cross Validation 5**

**Figure 5: Risk Stratification Over Cross Validation. Rate of extubation failure is given as a percentage for each bin.**

identified to increase the number of labeling functions and improve the performance of the model.

## 7.1 Auto Generation of Labeling Functions

In its current instantiation, the AutoWean system relies on the creation of labeling functions by clinicians, expert research, or data analysis. While this creates high quality labeling functions, it adds additional time and effort to an already labor intensive data collection phase. In the future, auto generation of labeling functions would be a powerful addition. For example, it should be possible to automatically learn decision trees with a depth of one. These decision trees should should be reminiscent of the basic labeling functions that are created for the AutoWean model. But, as many medical datasets are limited, it is important to verify that the labeling function not only makes sense from a data analytics aspect but also from the medical standpoint.

## 7.2 Labeling Functions

In this study, risk factors were selected based on the clinician's recommendations and domain-specific knowledge. Intuitively, incorporating more risk factors into our model would increase the number of labeling functions, enhancing the model performance. However, acquiring more risk predictors from ventilators was limited since ventilators from a specific vendor did not support certain variables. This made it hard for us to design more effective labeling functions as only part of the patients had data points for strong predictors.

Nonetheless, we can find additional variables from monitoring devices connected to the patient. Monitoring devices send HL7 messages that include variables such as vital sign measurements. Also, we can pull out patients' information from Admit, Discharge, Transfer (ADT) message that has just recently been started to be received by our data acquisition system. This would complement patients' demographic information, which our dataset lacks. Once appropriate risk factors are added to our dataset, the auto-generation of labeling functions can be used in combination with clinical verification to improve extubation risk estimation.

## 7.3 Open-Source Dataset

The AutoWean system is evaluated on the dataset collected from the Hospital of University of Pennsylvania instead of a widely used open-source dataset such as MIMIC-III [17]. Our initial research involved the MIMIC-III dataset but it yielded scarce entries after preprocessing and feature extraction. While analyzing the MIMIC-III dataset, we identified numerous erroneous values and timestamps, as well as high missing degrees for important features. In contrast, our collected dataset contains a sufficient number of extubations that have both accurate and high resolution time-series data. Although MIMIC-III could not be used for our AutoWean system, the model could be generalized for other medical datasets if they have an adequate number of data points for the labeling functions used in the AutoWean model.

# 8 CONCLUSION

In this paper, we presented AutoWean, a model that estimates extubation outcomes and provides risk stratification based on the labeling functions distinguishing between high-risk and low-risk patients. We evaluated 827 extubations identified from 494 patients who were ventilated in the ICUs across the Hospital of University of Pennsylvania from 2021 to 2022. The risk level is partitioned into five categories where a higher bin indicates higher risk. The result shows that extubations placed in high, med/high, and medium risk levels are prognostic of failures and successes in low, low/med risks. This meets our goal of aiding clinical decisions via suggesting extubation timing decisions through a feedback system with high confidence.

## ACKNOWLEDGMENTS

This research was supported in part by the U.S. National Science Foundation under grant 1915398 and the National Institute of Health under grant 1R01EB029767-01.

## REFERENCES

[1] Inmaculada Alía and Andrés Esteban. "Weaning from mechanical ventilation". In: *Critical care (London, England)* 4 (Feb. 2000), pp. 72–80. DOI: 10.1186/cc660.

[2] *Artificial Intelligence in healthcare market size report, 2030.* URL: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market.

[3] Niharika Bhardwaj et al. "The Impact of Big Data on Chronic Disease Management". In: *The Health Care Manager* 37 (Dec. 2017), p. 1. DOI: 10.1097/HCM.0000000000000194.

[4] Jean-Michel Boles et al. "Weaning from mechanical ventilation". In: *European Respiratory Journal* 29.5 (2007), pp. 1033–1056.

[5] Karen Burns et al. "Frequency of Screening and SBT Technique Trial - North American Weaning Collaboration (FAST-NAWC): A protocol for a multicenter, factorial randomized trial". In: *Trials* 20 (Oct. 2019). DOI: 10.1186/s13063-019-3641-8.

[6] Tingting Chen et al. "Prediction of extubation failure for intensive care unit patients using light gradient boosting machine". In: *IEEE Access* 7 (2019), pp. 150960–150968.

[7] Hyonyoung Choi et al. "VitalCore: Analytics and Support Dashboard for Medical Device Integration". In: *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE).* 2021, pp. 82–86. DOI: 10.1109/CHASE52844.2021.00016.

[8] Davide Cirillo and Alfonso Valencia. "Big data analytics for personalized medicine". In: *Current Opinion in Biotechnology* 58 (2019). Systems Biology - Nanobiotechnology, pp. 161–167. ISSN: 0958-1669. DOI: https://doi.org/10.1016/j.copbio.2019.03.004.

[9] Jonathan Cohen et al. "Prediction of extubation outcome: A randomised, controlled trial with automatic tube compensation vs. pressure support ventilation". In: *Critical care (London, England)* 13 (Mar. 2009), R21. DOI: 10.1186/cc7724.

[10] Thomas Davenport and Ravi Kalakota. "The potential for artificial intelligence in healthcare". In: *Future Hospital Journal* 6 (June 2019), pp. 94–98. DOI: 10.7861/futurehosp.6-2-94.

[11] William Doorn et al. "A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis". In: *PLOS ONE* 16 (Jan. 2021), e0245157. DOI: 10.1371/journal.pone.0245157.

[12] Alexandre Fabregat et al. "A Machine Learning decision-making tool for extubation in Intensive Care Unit patients". In: *Computer Methods and Programs in Biomedicine* 200 (2021), p. 105869.

[13] Fernando Frutos-Vivar et al. "Risk Factors for Extubation Failure in Patients Following a Successful Spontaneous Breathing Trial". In: *Chest* 130 (Jan. 2007), pp. 1664–71. DOI: 10.1378/chest.130.6.1664.

[14] David Haussler. *Probably approximately correct learning.* University of California, Santa Cruz, Computer Research Laboratory Santa, 1990.

[15] *Health Level Seven international.* URL: http://www.hl7.org/about/index.cfm?ref=common.

[16] Mohammad Hossein Jarrahi. "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making". In: *Business Horizons* 61.4 (2018), pp. 577–586. ISSN: 0007-6813. DOI: https://doi.org/10.1016/j.bushor.2018.03.007.

[17] Alistair Johnson, Tom Pollard, and Roger Mark. *Mimic-III clinical database demo.* Apr. 2019. URL: https://physionet.org/content/mimiciii-demo/1.4/.

[18] Shiho Kino et al. "A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects". In: *SSM - Population Health* 15 (June 2021), p. 100836. DOI: 10.1016/j.ssmph.2021.100836.

[19] Insup Lee et al. "Challenges and research directions in medical cyber–physical systems". In: *Proceedings of the IEEE* 100.1 (2011), pp. 75–90.

[20] Joon Lee. "Is Artificial Intelligence Better Than Human Clinicians in Predicting Patient Outcomes?" In: *Journal of Medical Internet Research* 22 (Aug. 2020), e19918. DOI: 10.2196/19918.

[21] Sean Lee et al. "Repeat Intubations Are Associated With More Procedural Complications in Multiply Intubated Patients". In: vol. 41. Dec. 2013. DOI: 10.1097/01.ccm.0000439504.28041.2d.

[22] Shuo Li et al. "PAC-Wrap: Semi-Supervised PAC Anomaly Detection". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, pp. 945–955. ISBN: 9781450393850. DOI: 10.1145/3534678.3539408. URL: https://doi.org/10.1145/3534678.3539408.

[23] John F McConville and John P Kress. "Weaning patients from the ventilator". In: *New England Journal of Medicine* 367.23 (2012), pp. 2233–2239.

[24] Babak Mokhlesi et al. "Predicting extubation failure after successful completion of a spontaneous breathing trial". In: *Respiratory care* 52 (Dec. 2007), pp. 1710–7.

[25] Jerome Osheroff et al. "A Roadmap for National Action on Clinical Decision Support". In: *Journal of the American Medical Informatics Association : JAMIA* 14 (Jan. 2007), pp. 141–5. DOI: 10.1197/jamia.M2334.

[26] Roberta Pastorino et al. "Benefits and challenges of Big Data in healthcare: an overview of the European initiatives". In: *European journal of public health* 29 (Oct. 2019), pp. 23–27. DOI: 10.1093/eurpub/ckz168.

[27] Lara Pisani, Nadia Corcione, and Stefano Nava. "Management of acute hypercapnic respiratory failure". In: *Current opinion in critical care* 22.1 (2016), pp. 45–52.

[28] Robert C Rothaar and Scott K Epstein. "Extubation failure: magnitude of the problem, impact on outcomes, and prevention". In: *Current opinion in critical care* 9.1 (2003), pp. 59–66.

[29] Christopher Seymour et al. "Minute Ventilation Recovery Time Measured Using a New, Simplified Methodology Predicts Extubation Outcome". In: *Journal of intensive care medicine* 23 (Jan. 2008), pp. 52–60. DOI: 10.1177/0885066607310302.

[30] Aracely Silva-Cruz et al. "Risk factors for extubation failure in the intensive care unit". In: *Revista Brasileira de Terapia Intensiva* 30 (Oct. 2018). DOI: 10.5935/0103-507X.20180046.

[31] Ida Sim et al. "Clinical Decision Support Systems for the Practice of Evidence-based Medicine". In: *Journal of the American Medical Informatics Association* 8.6 (Nov. 2001), pp. 527–534. ISSN: 1067-5027. DOI: 10.1136/jamia.2001.0080527. eprint: https://academic.oup.com/jamia/article-pdf/8/6/527/2336998/8-6-527.pdf. URL: https://doi.org/10.1136/jamia.2001.0080527.

[32] Reed Sutton et al. "An overview of clinical decision support systems: benefits, risks, and strategies for success". In: 3 (Feb. 2020). DOI: 10.1038/s41746-020-0221-y.

[33] Arnaud Thille, Jean-Christophe Richard, and Laurent Brochard. "The Decision to Extubate in the Intensive Care Unit". In: *American journal of respiratory and critical care medicine* 187 (May 2013). DOI: 10.1164/rccm.201208-1523CI.

[34] Amanda Watson et al. "RT-ACL: Identification of High-Risk Youth Patients and their Most Significant Risk Factors to Reduce Anterior Cruciate Ligament Reinjury Risk". In: *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. 2021, pp. 35–45. DOI: 10.1109/CHASE52844.2021.00012.

[35] Hossam Zein et al. "Ventilator Weaning and Spontaneous Breathing Trials; an Educational Review". In: *EMERGENCY* 4 (Apr. 2016).