Predicting and Understanding Unexpected Respiratory Decompensation in Critical Care Using Sparse and Heterogeneous Clinical Data

Oliver Ren[1*], Alistair E. W. Johnson[1*†], Eric P. Lehman[2], Matthieu Komorowski[3], Jerome Aboab[1], Fengyi Tang[1], Zach Shahn[4], Daby Sow[4], Roger G. Mark[1], Li-wei H. Lehman[1†]
[1] Massachusetts Institute of Technology, Cambridge, MA
[2] Northeastern University, Boston, MA
[3] Imperial College London, London, UK
[4] Center for Comp. Health, IBM Research, Yorktown Heights, NY

***Abstract*—Hospital intensive care units (ICUs) care for severely ill patients, many of whom require some form of organ support. Clinicians in ICUs are often challenged with integrating large volumes of continuously recorded physiological and clinical data in order to diagnose and treat patients. In this work, we focus on developing interpretable models for predicting unexpected respiratory decompensation requiring intubation in ICU patients. Predicting need for intubation could have important implications for the patient and medical staff and potentially enable timely interventions for improved patient outcome. Using data from adult ICU patients from the Medical Information Mart for Intensive Care (MIMIC)-III database, we developed gradient boosting models for predicting intubation onset. In a cohort of 12,470 patients, of whom 1,067 were intubated (8.55%), we achieved an area under the receiver operating characteristic curve (AUROC) of 0.89, with 95% confidence interval (CI) 0.87 - 0.91, when predicting intubation 3 hours ahead of time, a significant increase ($p<0.001$) over the AUROC achieved using several baselines, including logistic regression (0.81, 95% CI 0.78 - 0.84) and neural networks (0.80, 95% CI 0.77 - 0.83]). Finally, we conducted feature importance analysis using gradient boosting and derived useful insights in understanding the relative importance of clinical vs. biological variables in predicting impending respiratory decompensation in ICUs.**

I. Introduction

Modern intensive care units (ICUs) provide continuous monitoring of critically-ill patients, collecting large volumes of clinical and physiological data, including vital-signs (heart rate and blood pressure), and laboratory measurements (such as chemistry, hematology, and arterial blood gases, etc). Clinicians in the ICUs face the challenges of interpreting large volumes of data for timely diagnosis and treatment of patients, many of whom require organ support such as dialysis, vasopressors, and mechanical ventilation (MV).

MV requires intubation, which is the insertion of a tube inside the trachea of a patient to support or replace respiration. In an ICU setting, patients are mechanically ventilated for a number of conditions such as pneumonia, pulmonary edema, chest trauma, or reduced consciousness. MV maintains adequate oxygenation, reduces the work of breathing, and could protect diaphragmatic muscle fibers from the initial inflammatory insult [1]. Predicting the need for MV could potentially improve patient prognostication and, at an organizational level, allow better level of staffing since intubated patients in general require a 1:1 nurse to patient ratio.

Assessing the need for intubation currently relies on the expertise of medical staff, who interpret measurements from a wide range of variables, including both clinical (e.g. vital signs, urine output, and co-morbidities) and biological measurements (i.e. requiring blood samples). *Clinical* variables can be directly obtained by observing patients (e.g. vital signs). *Biological* measurements, on the other hand, are obtained via lab tests that require patient blood samples. Additionally, the results take time to develop, limiting the speed with which decisions based on biological measurements can be made; yet in many areas, they are essential in patient diagnoses and treatment. While clinicians use both clinical and biological variables in their medical decision process, studies that compare the relative importance of these variables have been sparse.

In this work, we use machine learning techniques to predict urgent need for intubation in intensive care units' patients. We aim to develop predictive models which are accurate, interpretable, and robust in the presence of sparse data samples. We trained gradient boosting models to predict MV onsets, and used the average reduction in Gini impurity across all trees to calculate feature importance in a multivariable setting. We compared the predictive performance of our approach with several baseline algorithms, including logistic regression, and neural networks using data from 12,470 adult patients from the Medical Information Mart for Intensive Care (MIMIC)-III database [2]. We conducted feature importance analysis to understand the relative importance of clinical vs. biological variables in predicting impending respiratory decompensation at different time points prior to the event onset.

II. Related Work

Several recent works demonstrated the effectiveness of using electronic health records for diagnosis and outcome

* Co-first authors.
† Corresponding authors - aewj@mit.edu, lilehman@mit.edu.

prediction [3], [4], [5], [6]. For example. Che et al., Lipton et al., and Razavian et al. predicted patient mortality, diagnoses, and disease onset respectively [4], [5], [6]. Prior work by Moss et al. used multivariable logistic regression to identify observable physiological signatures of respiratory failure that lead to unplanned intubation [7]. Suresh et al. used neural networks, specifically LSTM based models, to predict five different intervention tasks, including the onset of MV [8]. They approximated the relative importance of individual features by successively holding out individual features and evaluating the impact on the performance of their model [8].

In contrast, our study used gradient boosting to develop models for predicting unexpected intubation onset, and used the average reduction in Gini impurity across all trees to calculate feature importance. In addition, our choice of gradient boosting models has the added advantage of being robust against missing data.

Several previous studies on intubation prediction have been conducted. Some of these studies focus specifically on prolonged intubation in rather specific populations. Sharma et al. focused on predicting which patients would require intubation exceeding 48 hours after having cardiac surgery [9]. Figueroa-casas et al. identified three different regression models for predicting within the first two days whether a patient would require prolonged intubation [10]. A final study conducted by Walgaard et al. focused on identifying which factors were most predictive of prolonged MV in patients with Guillain-Barre syndrome [11].

In addition, there are a number of studies that aim to leverage machine learning techniques to help predict the outcome of weaning a patient off of MV. Mueller et al. used neural networks, support vector machines, Bayesian classifiers, decision trees, and logistic regression to find the classifier best suited for predicting the outcome of weaning an infant off MV [12]. Kuo et al. used a similar neural network approach to predict successful extubation, the removal of the inserted tube, in medical ICU patients who were mechanically intubated [13]. Prasad et al. built on that research and used reinforcement learning to come up with a strategy for weaning patients off of MV in the ICU [14].

In comparison, our paper focuses on a more general ICU population. Instead of focusing on prolonged intubation, our models predict whether a patient will need intubation regardless of the length of the treatment. Furthermore, preexisting models used fixed window sizes to predict intubation events [7], [8], whereas our research used varying data windows to create different predictive models. This allowed us to explore the effects of data window length on prediction accuracy and feature importance.

III. Methods

A. Dataset Development and Description

Data for this study was extracted from the Medical Information Mart for Intensive Care (MIMIC)-III database [2]. MIMIC-III is a large, publicly available database containing data for over 40,000 patients admitted to ICUs at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, USA. MIMIC-III contains high resolution data including time-stamped vital signs, laboratory values, treatment indicators, and billing codes.

1) Cohort: We used data from adult patients from MIMIC III in this study (age greater than 16 at admission). Next, for patients with multiple hospital or ICU stays, we only considered the first ICU stay for the first hospital stay. We excluded patients admitted under surgical service because surgical patients are frequently intubated due to anesthesia rather than respiratory failure. Patients who were intubated on admission were excluded, as were patients who were intubated or discharged from the ICU within 27 hours from admission. Finally, we removed patients who requested a withdrawal of care within the first 27 hours as these patients will not be intubated despite respiratory decompensation.

2) Data extraction: For each patient in our cohort, we extracted data from a window of size W hours, located L hours before a given event time. Specifically, we extracted a number of features, (see Table I), from the data window $[t_e - W - L, t_e - L]$, where t_e represents the event time. For intubated patients, the event time was the time of intubation. For non-intubated patients, the event time was a random time after the 27th hour during their ICU stay. For non-intubated patients with a code status change (e.g. changed to do not resuscitate or do not intubate), we ensured that the event time was before the code status change. For all patients, we evaluated and extracted features from data windows with window sizes of $W \in \{8, 12, 16, 20, 24\}$ hours and lead times of $L \in \{3, 6, 9, 12, 15\}$ hours.

Among these features were (i) vital signs, (ii) blood gas measurements, and (iii) laboratory measurements. Additional features included the total urine output over the data window and the presence of two comorbidities (congestive heart failure and pulmonary circulation disorders). In Table I, blood gas measurements and laboratory measurements, which require the use of a laboratory, are labeled as *biological*. The remaining variables, including vital signs, and comorbidities are labeled as *clinical* features as they are readily available and can be easily measured by nurses and clinicians at the bedside.

B. Model Development and Evaluation

We built four types of models for predicting MV onsets: (1) a linear model using logistic regression, (2) a neural network model, using a simple two layer feed forward network with ReLU activations, (3) a joint denoising autoencoder [15] and neural network model (NN/DAE), and (4) gradient boosting.

Gradient boosting was implemented using the *xgboost* package v0.60, while logistic regression was implemented using the *scikit-learn* package v0.18 [16], [17]. After grid searching (using the training data) through [100, 200, 300, 400, 500] trees, learning rate of [0.1, 0.2, 0.3] and a max depth of [3, 4, 5, 6], the following hyperparameters were used for gradient boosting: 100 trees, learning rate of 0.1, and a max depth of 4.

TABLE I
LIST OF ALL THE FEATURES EXTRACTED FOR EACH PATIENT.

Feature type	Feature extracted	Variables
Vital signs (clinical)	First, Last	Age, Gender, Heart rate, Systolic/Diastolic/Mean blood pressure, Respiratory rate, Temperature, Peripheral Oxygen Saturation, Glasgow Coma Scale
Blood gas, lab measurements (biological)	Last	Partial pressure of oxygen, Partial pressure of carbon dioxide, pH, Base excess, Bicarbonate, Total carbon dioxide concentration, Hematocrit, Hemoglobin, Carboxyhemoglobin, Methemoglobin, Chloride, Calcium, Temperature, Potassium, Sodium, Lactate, Anion gap, Albumin, Immature band forms, Bicarbonate, Bilirubin, Creatinine, Chloride, Glucose, Hematocrit, Hemoglobin, Platelet, Potassium, Partial thromboplastin time, International Normalized Ratio, Prothrombin time, Blood urea nitrogen, White blood cell count
Additional (clinical)	Sum	Urine output
Comorbidities (clinical)	Yes/No	Presence of congestive heart failure, Presence of pulmonary circulation disorders

TABLE II
COHORT CHARACTERISTICS (MEDIAN AND IQR SHOWN).

	No MV	MV
Patient count	11403	1067
Age	65 (51, 78)	67 (55, 77)
Male	6255 (54.85%)	636 (59.61%)
Resp rate (breath/min)	19 (16, 22)	22 (17, 27)
WBC ($10^3/mm^3$)	9 (6.6, 12.1)	10.6 (7.8, 14.9)
Arterial pO2 (mmHg)	82.5 (63, 108)	109 (73, 296.25)
GCS	15 (15, 15)	15 (14, 15)
Heart Rate (bpm)	81 (70, 94)	90 (76, 105)
SpO2 (%)	97 (95, 98)	96 (94, 98)
Temperature (°C)	36.7 (36.3, 37.0)	36.7 (36.2, 37.2)
Platelet count ($10^9/L$)	196 (142, 262)	183 (121.5, 264)
Creatinine (mg/dL)	1 (0.7, 1.5)	1.1 (0.7, 1.7)
BUN (mg/dL)	19 (12, 33)	24 (16, 40)
Glucose (mg/dL)	114 (96, 141)	130 (108, 162.25)
Urine output (mL)	1755 (1090, 2680)	1413.5 (834.5, 2178.5)

The NN/DAE model used a denoising autoencoder (DAE) to learn a low-dimensional, non-linear embedding of the observations; this low-dimensional representation was then used as input to a feed-forward neural net for outcome prediction. The DAE consists of a two-layer architecture for encoding and decoding respectively; ReLU activation was used for the first layer, and a sigmoid activation was used for a 16-dimensional middle-layer, with added Gaussian noise. In the NN/DAE model, output from the 16-dimensional middle layer of the DAE was fed into a feed-forward neural network, with a ReLU layer, followed by a dropout and sigmoid layer for prediction. Grid search was performed through [64, 128] hidden units for the ReLU layers, and a dropout of [0.2, 0.4, 0.5]. A learning rate of 0.001 was used.

All neural network based models were optimized using the Adam optimizer [18]. The NN model was optimized over 200 epochs. After grid searching through [128, 256, 512] hidden units per layer and a learning rate of [0.0001, 0.0002, 0.0003], the following hyperparameters were used for the final neural net model: 128 hidden units per layer and a learning rate of 0.0001. For logistic regression and neural nets, we imputed missing data using the population average of the training set. For gradient boosting, the algorithm automatically assigned contributions to missing data within trees in order to improve model fit.

We split the dataset into training (70%) and test sets (30%), holding out the test set to use only for the final evaluation. We used 10-fold cross-validation on the training set to train our models, and performed grid search for the best hyperparameters for both gradient boosting and neural network based models. For all our models, performance was evaluated using the area under the receiver operator characteristic curve (AUROC). Confidence intervals and comparisons for area under the receiver operating curves (AUCs) for the test set were based on the method described in [19]. Two-sided p values less than 0.05 were considered statistically significant. In tests that involve multiple comparisons, test of statistical significance was based on p values after correcting for multiple comparisons using FDR (false discovery rate) [20].

We used the average reduction in Gini impurity across all trees from gradient boosting models as a measure of relative feature importance of clinical vs. biological variables in predicting impending respiratory decompensation. To uncover how the importance of an individual feature varied alongside the lead size, we created a ranking of the top 10 features for gradient boosting models with $W = 24$ as L varied from 3 to 15 hours.

IV. RESULTS

A. Dataset Description

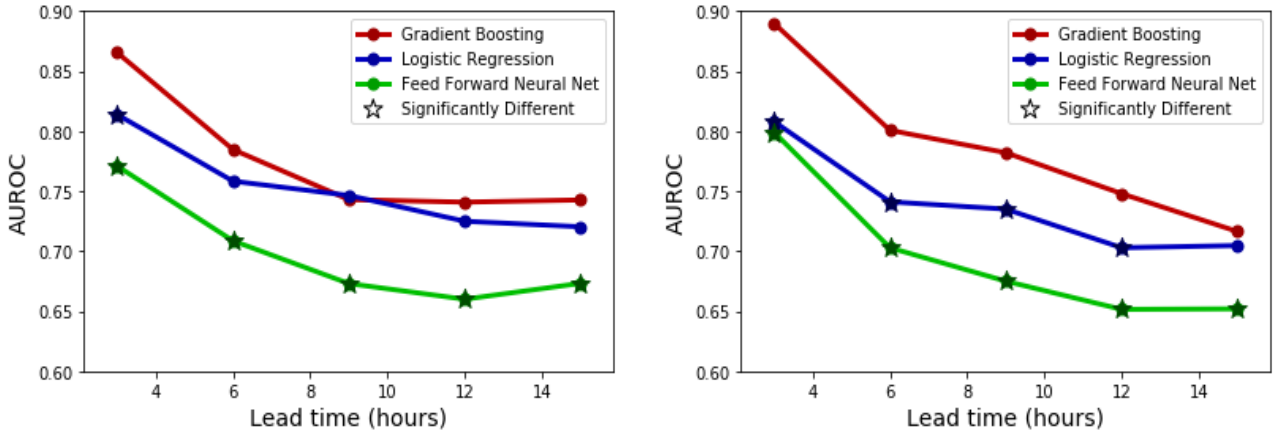
Application of exclusion criteria resulted in 12,470 patients, of whom 1,067 were intubated after the first day. Table II compares patients with and without ventilation, and displays for both groups the population median and interquartile range (IQR) of several variables. For both groups, the data was drawn from a window of 24 hours, ($W = 24$). For the MV group, the data window was 3 hours prior to the onset ($L = 3$). For biological variables, only the last measurement in the data window was considered. The median age of patients was 65, and 6,891 were male (55.3%). When comparing patients with and without intubation, intubated patients had higher median respiratory rate, white blood cell count, heart rate, and blood urea nitrogen in addition to a lower platelet count and urine output.

B. Performance as a Function of Lead Time and Observation Window Size

Table III compares prediction performance of gradient boosting with several baseline models at different window

TABLE III
(ALL FEATURES) PERFORMANCE AS DETERMINED BY AUROC (AND 95%CI) FOR VARYING WINDOW SIZES (W) AND LEAD TIMES (L). THE P-VALUES FOR COMPARISONS BETWEEN THE LOGISTIC REGRESSION AND THE GRADIENT BOOSTING MODELS ARE ALSO LISTED. SIGNIFICANTLY DIFFERENT LOGISTIC REGRESSION AND GRADIENT BOOSTING MODELS ARE MARKED WITH A *.

Data set	Neural Net	Neural Net/Autoencoder	Logistic Regression	Gradient Boosting	p-value
$W = 24$, $L = 3$	0.77 (0.74, 0.80)	0.77 (0.74, 0.80)	0.81* (0.79, 0.84)	0.87* (0.84, 0.90)	<0.0001
$W = 8$, $L = 3$	0.80 (0.77, 0.83)	0.78 (0.75, 0.81)	0.81* (0.78, 0.84)	0.89* (0.87, 0.91)	<0.0001
$W = 24$, $L = 15$	0.67 (0.64, 0.71)	0.70 (0.66, 0.73)	0.72 (0.69, 0.75)	0.74 (0.71, 0.78)	0.0919
$W = 8$, $L = 15$	0.65 (0.62, 0.69)	0.69 (0.66, 0.72)	0.71 (0.67, 0.74)	0.72 (0.68, 0.75)	0.3911



(a) Data window of 24 hours.

(b) Data window of 8 hours.

Fig. 1. (All features) Performance of the models as the lead time varied from 3 to 15 hours. Models that are starred (*) are significantly different than the gradient boosting models using the same window size and lead time. Data window size was fixed at 24 hours (Panel a) and 8 hours (Panel b) respectively.

TABLE IV
(ALL FEATURES) PERFORMANCE OF GRADIENT BOOSTING MODELS MEASURED BY SENSITIVITY AND SPECIFICITY FOR VARYING WINDOW SIZES (W) AND LEAD TIMES (L).

Data set	Sensitivity	Specificity
$W = 24$, $L = 3$	0.782	0.794
$W = 8$, $L = 3$	0.791	0.809
$W = 24$, $L = 15$	0.707	0.694
$W = 8$, $L = 15$	0.723	0.604

sizes and lead times in terms of their AUROCs and 95% CIs. The sensitivity and specificity of the gradient boosting models are presented in Table IV. Gradient boosting had the highest AUROC of 0.89 (0.87, 0.91) when $W = 8$ and $L = 3$ hours, significantly outperforming logistic regression when the lead time was 3 hours. When the lead time was increased to 15 hours (i.e. prediction was performed further away from the event), the performance for all models decreased and there were no statistically significant differences between gradient boosting and logistic regression. Gradient boosting outperformed neural networks based approaches (p-values $<$ 0.0001) in all window and lead time settings presented in Table III.

Figure 1a and b compared the predictive performance of gradient boosting, logistic regression and neural networks models as a function of the lead time. Figures 1a and 1b plot the performance of the models when the window size was fixed at 24 and 8 hours respectively as the lead time varied from

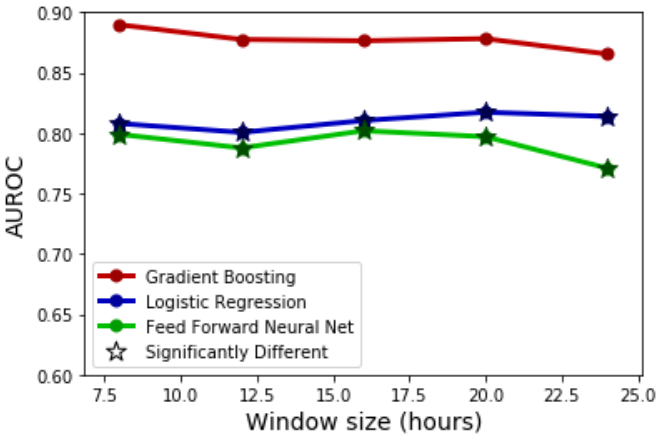
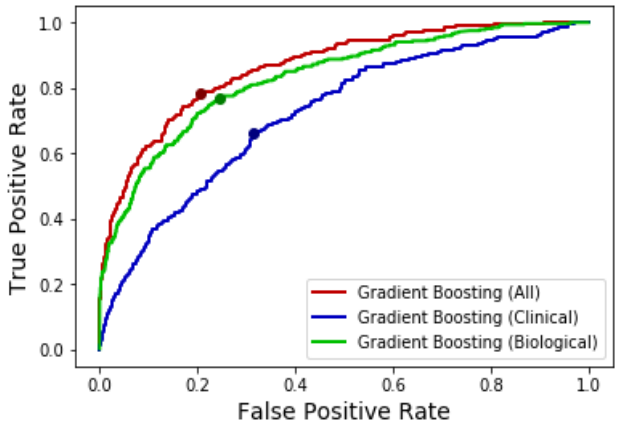


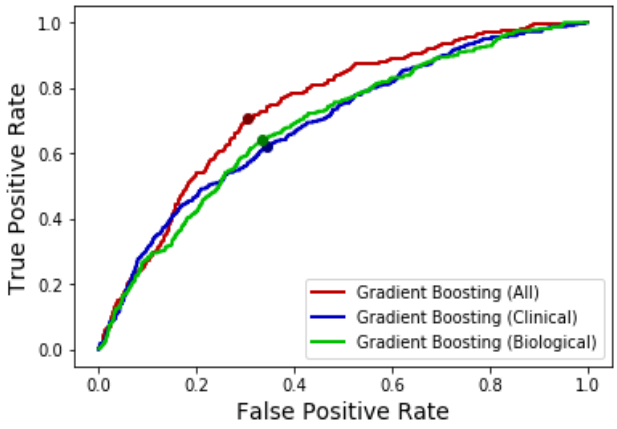
Fig. 2. (All features) Performance of gradient boosting and logistic regression models as the window size varied from 8 hours to 24 hours. The lead time was fixed to 3 hours. AUROCs that are starred indicate significant differences with the gradient boosting model using the same window size.

3 to 15 hours (i.e. as the prediction was performed further from the time of the intubation event). Performance of all three models decreased as the lead time increased. In Figure 1a and 1b, logistic regression and neural network models that were significantly different than gradient boosting models after adjusting for FDR were marked with a *.

Figure 2 plots the performance of models when using a fixed lead time of 3 hours, while the observation window

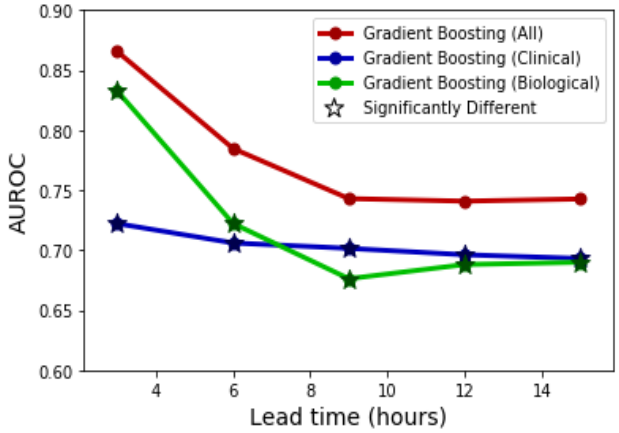


(a) Gradient boosting models $W = 24$, $L = 3$.

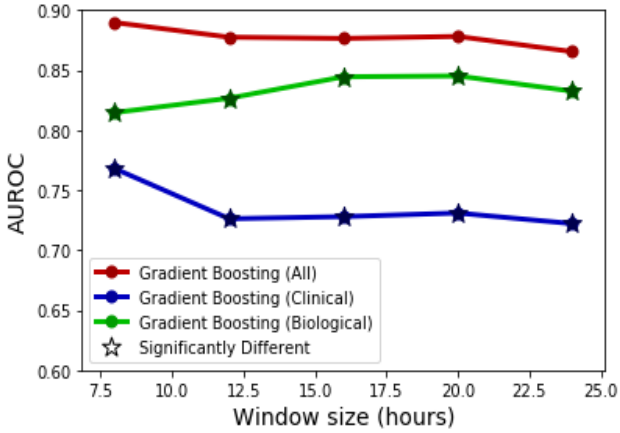


(b) Gradient boosting models $W = 24$, $L = 15$.

Fig. 3. ROC curves of gradient boosting when using all the features, when using only clinical features, and when using only biological features. The plotted points are the respective points closest to having a true positive rate of 1 and a false negative rate of 0. Two different data windows were considered: $W = 24$, $L = 3$ (Panel a), $W = 24$, $L = 15$ (Panel b).



(a) Fixed window size of 24 hours.



(b) Fixed lead time of 3 hours.

Fig. 4. Performance of gradient boosting when using all the features vs. when using clinical or biological only features. AUROCs that are starred are significantly different than the gradient boosting model with all the features. Two different data windows were considered: window size was fixed at 24 hours (Panel a) and lead time was fixed at 3 hours (Panel b).

TABLE V
PERFORMANCE OF GRADIENT BOOSTING MODELS AS DETERMINED BY AUROC (AND 95%CI) FOR VARYING WINDOW SIZES (W) AND LEAD TIMES (L) USING CLINICAL ONLY VS. BIOLOGICAL ONLY MEASUREMENTS. P-VALUES ARE FROM COMPARING THE TWO MODELS.

Data set	Clinical Only	Biological Only	p-value
$W = 24$, $L = 3$	0.72 (0.69, 0.76)	0.83 (0.81, 0.86)	<0.0001
$W = 8$, $L = 3$	0.77 (0.74, 0.80)	0.82 (0.79, 0.84)	0.0205
$W = 24$, $L = 15$	0.69 (0.66, 0.73)	0.69 (0.66, 0.72)	0.8785
$W = 8$, $L = 15$	0.70 (0.67, 0.73)	0.63 (0.60, 0.67)	<0.01

size varied from 8 to 24 hours. For both logistic regression and gradient boosting, the models did not have significantly different performance across all window sizes (individual comparisons, $p > 0.05$).

C. Comparing Importance of Clinical vs. Biological Variables

We compared the performance of models using biological vs. clinical features to approximate how much of the overall gradient boosting model's performance could be attributed to each of these subsets individually. Figure 3a and 3b compare the ROC curves for gradient boosting models which consider all the features vs. clinical or biological figures only. Performance of the gradient boosting models when considering either clinical or biological only features for $W = 24$ or 8 hours and $L = 3$ or 15 hours are shown in Table V. For the clinical only models, the AUROCs dropped drastically, from 0.77 to 0.70, when the lead time was varied from 3 to 15 hours ($W = 8$). However, when we increased the window size to 24 hours, as Figure 4a shows, the AUROC stayed relatively constant as we vary the lead time.

However, for the biological only models, the AUROCs

dropped drastically when varying lead times from 3 to 15 hours for both a fixed window size of 8 hours, from 0.82 to 0.63, and a fixed window size of 24 hours, from 0.83 to 0.69. When comparing the clinical only models to the biological only models, the clinical only models outperformed the biological only models when the lead time is large ($L = 15$), despite under-performing when the lead time is small ($L = 3$).

In Figure 4b, we note that when we fixed the lead time to 3 hours, the performance gap between biological vs. clinical only models increased as the data window increased. This is potentially because as we increased the data window of clinical only models, the first measurements become less representative of the patients' state prior to intubation, thus bringing down the performance. For the biological only models, as the data windows increased, the performance also increased, potentially due to the fact that more measurements were included in the analysis with a large observation window size.

Unlike the models that considered other subsets of features, the gradient boosting models that only used clinical features did not have changes in performance when lead times change. The flat slope of the gradient boosting AUROC curve for clinical only gradient boosting models as lead time changes in Figure 4a implies that patient state using clinical variables is a fairly consistent way to determine whether a patient will need intubation.

D. Feature Importance Analysis

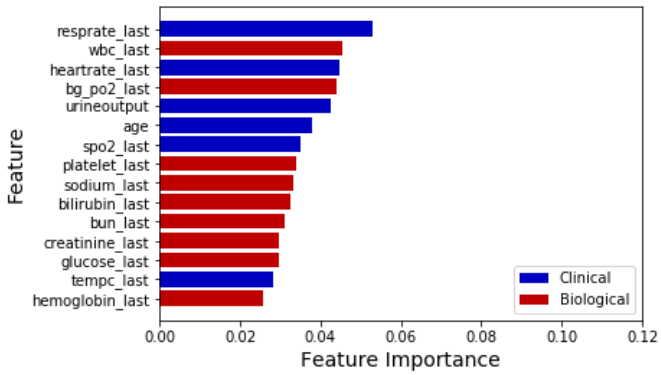


Fig. 5. (All features) Top 15 most important features in the gradient boosting model with a window size of 24 and a lead time of 3 as ranked by the average reduction in Gini impurity across all trees for a given feature.

Figure 5 shows the most important features from gradient boosting models created using all the features. Interestingly, five of the top seven most important features are clinical ones. This is surprising because results from the previous section indicated that the biological only models outperformed the clinical only models for all window sizes with a fixed lead time of 3 hours. The superior performance from the biological only models might be due to the fact that there were more biological features than clinical ones. As a result, although the biological features overall achieved higher predictive performance, analysis on relative importance of individual features revealed that several clinical features played a more important role in predicting respiratory decompensation.

We present the most important features of the clinical and biological only models in Figure 6a and 6b respectively ($W = 24$ hours, $L = 3$). Note that many of the top features were non-respiratory features, such as urine output, age, heart rate, temperature, white blood count, platelet count.

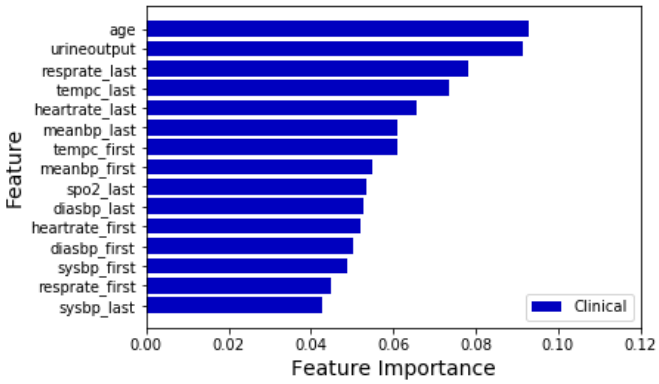
A ranking of the most important features across different lead times but a fixed window size of 24 hours for gradient boosting models using all the features is shown in Figure 7. The top 10 features were chosen independently for each model, and the resultant union of these features (16 in total) is presented. Some features such as last respiratory rate, and urine output became more important as we moved closer to the intubation event. Some features, such as last platelet count and first heart rate became less important the closer we were to the intubation event. Finally, some features such as age and last bilirubin had a relatively consistent importance as lead times decreased (i.e. closer to intubation). While we observed interesting trends in feature importance, we caution against drawing hard conclusions based on the relative change, as feature importance from different gradient boosting models may not be directly comparable.

V. Discussion & Conclusions

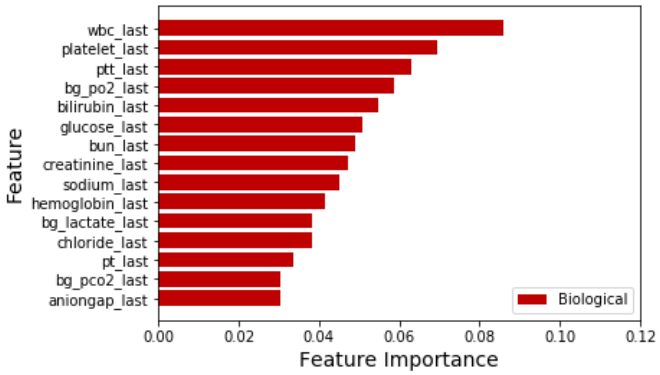
In this paper, we show that it is possible to develop an automated early warning system to alert clinicians of patients with impending respiratory failure, while providing interpretable results to elucidate the importance of individual features prior to the events. We demonstrated that gradient boosting predictive models can be the basis of such a system as it achieved a reasonable predictive performance while maintaining interpretability.

While we have demonstrated that gradient boosting performed better than several baseline models, including logistic regression and neural networks, when a simple missing data imputation technique was used for these baselines, we note that several recent techniques have been proposed to explicitly encode features from missing data to improve prediction performance [4], [21]. We leave comparisons with more advanced imputation or missing-data feature engineering techniques for future work.

We showed that the task of predicting MV onsets becomes drastically more difficult the farther away a patient is from the intubation event and that clinical and non-respiratory features are important during prediction. This has important implications to enable timely interventions, for patient prognostication and staffing needs. When observing the most important features for the gradient boosting model that considers all features, we noted a large prevalence of clinical features. This is an encouraging insight because clinical features are readily available at the bedside. In particular, being able to more heavily weight clinical features enables a more timely intubation decision-making process as clinical features are more readily available at the bedside. When examining the specific clinical and biological features that make up the most



(a) (Clinical features only) Gradient boosting model $W = 24$, $L = 3$.



(b) (Biological features only) Gradient boosting model $W = 24$, $L = 3$.

Fig. 6. Top 15 most important features in the gradient boosting model with a window size of 24 and a lead time of 3 as ranked by the average reduction in Gini impurity across all trees for a given feature. Only clinical features were considered in Panel a and only biological features were considered in Panel b.

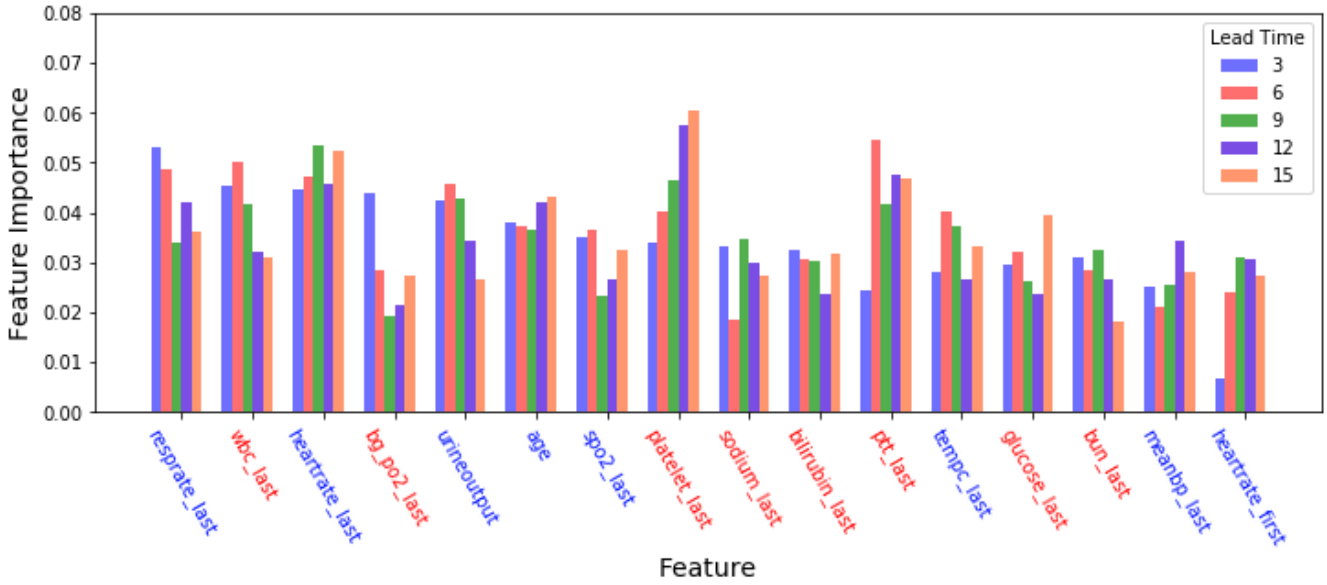


Fig. 7. (All features) Change in the top 10 most important features for gradient boosting models with a constant window size ($W = 24$) over different lead times ($L = \{3, 6, 9, 12, 15\}$). Blue feature labels are clinical and red feature labels are biological. Feature importance is determined by average reduction in Gini impurity across all trees for a given feature.

important features lists, we note that many of the top features are non-respiratory. This is shown explicitly by Figures 6a and 6b, which show the most important features of a gradient boosting model that considers only clinical features and one that considers only biological features respectively. For example, Figure 6a shows that urine output is more important and mean blood pressure and heart rate are almost as important as respiratory rate when predicting the need for intubation. This may reflect situations such as renal failure or cardiac failure leading to fluid accumulation and pulmonary edema, a condition that may require MV.

We note the following limitations to our models and methodology. There is the issue of 1) missing information about the true clinical states. There are many indications for intubation and MV, some of which are not or poorly translated in the numerical data, or not easily retrievable (for example present only in the clinicians' notes): airway obstruction, severe agitation or delirium, pain control, complication requiring surgery, etc. Another issue is 2) causality leakage, which is especially possible if the lead time is short. For example, if a patient is sedated to be intubated, the GCS will drop a few minutes before the intubation. Predicting the need for intubation based on recent GCS could represent a case of causality leakage. Finally, there is the issue of 3) selected windows being based off of intubation time - which is acausal as it assumes we know when a patient is intubated - nevertheless it allows us to infer about factors related to intubation. Therefore, practical applications of the models described in this paper in a clinical setting would require a new framework for evaluation. These are just some factors that

can negatively affect the performance and generalizability of our models.

Finally, while early onset of MV could be beneficial, MV can be a double-edged sword with potential serious adverse effects such as hemodynamic impairment, ventilation induced injury and ventilator acquired pneumonia. To date, there is no clinical or experimental data defining neither optimal timing nor best criteria for initiation of MV. As there is no scientific rationale, there is a great variability in actual practice which is mainly based on clinicians convictions and experience. A recent 2016 global survey conducted by de Montmollin et al. [22] reported a lack of consensus in initiation strategies for MV. For some conditions such as respiratory failure, there was a general consensus that patients should be intubated, but for other conditions such as cardiovascular failure, no overwhelming consensus was reached. Ebihara et al. showed that MV has the positive benefit of protecting against diaphragm damage in septic patients [1]. However, in the context of critical care patients, Vassilakopoulos et al. conducted research supporting the contrary statement that MV can actually induce diaphragm damage [23]. Thus, an important area of future research is to determine the optimal initiation strategies for MV.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH) grant R01-EB017205, R01-EB001659 and R01GM104987 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIBIB or the NIH.

REFERENCES

[1] S. Ebihara, S. Hussain, G. Danialou, W.-K. Cho, S. Gottfried, and B. Petrof, "Mechanical ventilation protects against diaphragm injury in sepsis: interaction of oxidative and mechanical stresses," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, pp. 221–228, 2002.

[2] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. G. Mark, "MIMIC-III: a freely accessible critical care database," *Scientific Data*, vol. 3, no. 160035, 2016.

[3] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277–287, 2015.

[4] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8.6085, 2016.

[5] Z. C. Lipton, D. C. Kale, C. Elkan, and R. C. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," *ICLR*, 2016.

[6] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal laboratory tests," in *Proceedings of the 1st Machine Learning for Healthcare Conference*, vol. 56. PMLR, 2016, pp. 73–100.

[7] T. Moss, D. Lake, F. Calland, K. Enfield, J. Delos, K. Fairchild, and J. Moorman, "Signatures of subacute potentially catastrophic illness in the icu: Model development and validation," *Critical Care*, vol. 44, no. 9, pp. 1639–1648, 2016.

[8] H. Suresh, N. Hunt, A. E. W. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical intervention prediction and understanding using deep networks," *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 2017.

[9] V. Sharma, V. Rao, C. Manlhiot, A. Boruvka, S. Fremes, and M. Wasowicz, "A derived and validated score to predict prolonged mechanical ventilation in patients undergoing cardiac surgery," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 153, no. 1, pp. 108–115, 2017.

[10] J. B. Figueroa-Casas, A. K. Dwivedi, S. M. Connery, R. Quansah, L. Ellerbrook, and J. Galvis, "Predictive models of prolonged mechanical ventilation yield moderate accuracy," *Journal of Critical Care*, vol. 30, no. 3, pp. 502–505, 2015.

[11] C. Walgaard, H. F. Lingsma, P. A. van Doorn, M. van der Jagt, E. W. Steyerberg, and B. C. Jacobs, "Tracheostomy or not: Prediction of prolonged mechanical ventilation in guillainbarr syndrome," *Neurocritical Care*, vol. 26, no. 1, pp. 6–13, 2017.

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

[13] H. Kuo, H. Chiu, C. Lee, T. Chen, C. Chang, and M. 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.

[14] N. Prasad, L. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2017.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, 2010.

[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *arXiv preprint arXiv:1603.02754*, 2016.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[20] Y. Benjamini and H. Y., "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. 57, no. 1, pp. 289–300, 1995.

[21] Z. C. Lipton, D. C. Kale, and R. C. Wetzel, "Directly modeling missing data in sequences with rnns: Improved classification of clinical time series," *In Machine Learning for Healthcare Conference*, 2016.

[22] E. de Montmollin, J. Aboab, R. Ferrer, E. Azoulay, and D. Annane, "Criteria for initiation of invasive ventilation in septic shock : An international survey," *Journal of Critical Care*, vol. 31, pp. 54–57, 2016.

[23] T. Vassilakopoulos and B. Petrof, "Ventilator-induced diaphragmatic dysfunction," *American Journal of Respiratory and Critical Care Medicine*, vol. 169, pp. 336–341, 2004.