IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 15, 2019, accepted October 7, 2019, date of publication October 11, 2019, date of current version October 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2946980

Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine

TINGTING CHEN[1,2,*], JUN XU[3,*], HAOCHAO YING[1,2,*], XIAOJUN CHEN[2], RUIWEI FENG[1,2], XUELING FANG[3], HONGHAO GAO[4], AND JIAN WU[1,2]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[2]Real Doctor AI Research Centre, Zhejiang University, Hangzhou 310027, China
[3]Intensive Care Unit, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou 310027, China
[4]Computing Center, Shanghai University, Shanghai 200444, China

Corresponding authors: Xueling Fang (1191012@zju.edu.cn), Honghao Gao (gaohonghao@shu.edu.cn), and Jian Wu (wujian2000@zju.edu.cn)

*Tingting Chen, Jun Xu, and Haochao Ying are co-first authors.

This work was supported in part by the Subject of the Major Commissioned Project—"Research on China's Image in the Big Data" of Zhejiang Province's Social Science Planning Advantage Discipline—"Evaluation and Research on the Present Situation of China's Image" under Grant 16YSXK01ZD-2YB, in part by the Ministry of Education of China under Grant 2017PT18, in part by the Zhejiang University Education Foundation under Grant K18-511120-004, Grant K17-511120-017, and Grant K17-518051-021, in part by the Major Scientific Project of Zhejiang Laboratory under Grant 2018DG0ZX01, in part by the National Natural Science Foundation of China under Grant 61672453, and in part by the Key Laboratory of Medical Neurobiology of Zhejiang Province.

ABSTRACT Extubation failure is a complex and ongoing problem in the intensive care unit (ICU). It refers to the patients who require re-intubation after extubation (namely disconnection from mechanical ventilation). In these patients, extubation failure leads to severe risks associated with re-intubation and is associated with increased mortalities, longer stay in ICU and also higher health care costs. Many studies have been proposed to analyze the problem of extubation failure and identify possible factors or indices that may predict extubation failure. However, these studies used a small number of patients for extubation failure and limited their features to several vital signs or main characteristics. We argue that these are insufficient and less accurate for the prediction of extubation failure. In this paper, we analyze 3636 adult patient records in the MIMIC-III clinical database and apply the Light Gradient Boosting Machine (LightGBM) to predict extubation failure. Also, we perform feature importance analysis according to the result of LightGBM and interpret these features using SHapley Additive exPlanations (SHAP). Experimental results show that our LightGBM method is effective in predicting extubation failure and outperform other machine learning methods such as artificial neural network (ANN), logistic regression (LR) and support vector machine (SVM). The results of feature importance and SHAP analysis are also proved effective and accurate.

INDEX TERMS Extubation failure prediction, feature importance, light gradient boosting machine, shapley additive explanations.

I. INTRODUCTION

Mechanical ventilation is the primary way of breathing support for the patients in intensive care units (ICU). It is a life-saving medical procedure and can assist patients with acute respiratory difficulties. Studies have shown that around 50% of patients in ICU require mechanical ventilation [1], [2]. Extubation refers to the process of removing endotracheal tube from patients. It is the final step in liberating a patient from mechanical ventilation. In critical care, timely and effective extubation is an important goal [2].

The extubation process for a patient with mechanical ventilation should be performed as soon as possible. Once there is improvement noted within a few days or weeks, a clinical decision should be made to withdraw the endotracheal tube (i.e., extubation). Unnecessary delay and too early extubation both have adverse effects on patients [3], [4]. Delayed extubation (or long-term mechanical ventilation) is associated

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

with increased morbidities and even mortalities as well as a range of medical complications (e.g., ventilator-associated pneumonia, airway trauma) [5]–[7]. Extubation too early may cause extubation failure (EF). According to studies, up to 25% of patients fail the extubation and require re-intubation due to recurrence of respiratory insufficiency [8]. EF puts patients at acute risks associated with reintubation and increases health care costs and length of stay in ICU and also mortalities [9]–[11]. Therefore, considering the risks of delayed and too early extubation, it is significant to identify the ideal time point for extubation and reduce the patient's time on mechanical ventilation.

In clinical practice, the assessment for extubation readiness is mainly based on the outcomes of the spontaneous breathing trial (SBT), which contains T-tube breathing or low-level pressure support ventilation over at least 30 minutes [2]. In addition, the decision to extubate is also based on many other observations, such as blood gas results, ventilator settings, and clinical expertise. Despite advances in technology, the extubation decision making is still a complex and difficult task for clinicians [12]. Actually, there is no consensus on a standardized or objective weaning protocol [13], and it is controversial which parameters should be included in that protocol, even though the use of weaning protocols sometimes could be helpful [14]. Further, the actual applying of weaning protocol varies individually and even varies amongst different insititutions [13]. To help clinicians make more informed decision, an automated prediction tool for extubation is necessary.

Several prospective and retrospective studies have been proposed to identify important factors and indices that can predict extubation failture [15], [16]. The factors are ranging from demographic information (e.g., age, reason for intubation) [17], vital signs (e.g., heart rate, respiratory rate) [18], blood gas analysis (e.g., sodium, potassium, serum anion gap) [15], and pulmonary characteristics (e.g., duration of mechanical ventilation, tidal volume) [17], [19]. Saugel *et al.* [15] applyed statistic analysis on the medical records of 61 ICU patients and found that low pre-extubation serum anion gap values and low pre-extubation PaO_2/FiO_2 [1] ratio might be helpful. Chaparro and Giraldo [16] proposed a new extubation index based on the power of respiratory flow signal. Other indices, such as cardiorespiratory behavior [20] and breathing patterns [21] have been used for preterm neonate patients.

Recently, artificial neural network (ANN) has been applied by many researchers to address the problem of ventilator extubating. Gottschalk *et al.* [19] has utilized 4 variables (i.e., tidal volume [V_T], minute ventilation, breathing frequency, and maximum inspiratory pressure [$P_{I_{max}}$]) to train an ANN model. Kuo *et al.* [17] designed their ANN model according to 8 input variables, consisting of subjects' age, reasons for intubation, duration of mechanical ventilation, acute physiology and chronic health evaluation (APACHE) II scores, and breathing patterns obtained during a 30-min SBT. Their ANN model showned better discrimination than existing predictors, such as RSBI, $P_{I_{max}}$, $RSBI_1$, $RSBI_{30}$, and $\triangle RSBI_{30}$, [2] in predicting successful extubation. However, the data numbers for these ANN methods were small. Neural networks are more suitable for finding patterns in large sample data, and the training algorithm, back propagation (BP), for ANN also needs large numbers of data since the uncontrolled convergence speed and local optima problem of BP [22].

Mikhno and Ennett [18] applied machine learning techniques to the 179 neonate records in the MIMIC-II (Multi-Parameter Intelligent Monitoring of Intensive Care) clinical database [23] to locate features relevant to EF, and to develop a model for predicting EF in the neonatal intensive care unit. But they used logistic regression to screen all combinations of 3 features from a pool of 57 features (yielding 58,520 candidate models), and find 6 features for EF prediction from the top two models. We argue that this feature selection process is complicated and time-consuming.

Furthermore, all the methods above have a limitation, which is that they have a small number of patients for EF, and the numbers of EF patients are only 20 or 30. To improve the performance for the prediction of extubation outcome, studies with larger and more heterogeneous patient collectives are needed [15]. In addition, existing works have limited their features for prediction to at most a couple of key vital signs or main characteristics, which are not sufficient and less accurate for the extubation outcome analysis.

To these ends, in this paper, we propose an automated EF prediction and feature analysis model, which is based on the Light Gradient Boosting Machine (LightGBM). We extract and utilize 3636 adult patient records with nearly one hundred features from the MIMIC-III clinical database [24]. At the initial procedure, we perform the feature pre-processing, including missing value processing and feature correlation analysis. Then, we explore the use of LightGBM model to predict EF and analyze features that effect the extubation outcomes. The experimental results demonstrate the superiority of our method. The main contributions of our work are summarized as follows:

1) We employ a LightGBM model for EF prediction based on large patient records;
2) We incorporate a large number of possible features to perform a comprehensive analysis for EF prediction, initially with 92 features;
3) We perform the feature importance analysis and visualize important features with the SHAP method.

The remaining of this paper is organized as follows: Section II describes the data and methods used in this study. In Section III, we present the results, and Section IV gives

[1] PaO_2 and FiO_2 are the arterial partial pressure of oxygen and fraction of inspired oxygen respectively.

[2] RSBI = rapid shallow breathing index. $RSBI_1$ and $RSBI_{30}$ are RSBI at 1 min and 30 min in an SBT, respectively. $\triangle RSBI_{30}$ refers to the absolute percentage change in RSBI from 1 to 30 min in an SBT.

TABLE 1. **Preliminary features listed by type.**

Feature Types	Features
Demographic information	Age, Gender, Weight, Height, Weight loss, Chronic cardiac insufficiency, Arrhythmia, Valvular disease, Pulmonary heart disease, Peripheral vascular, Hypertension, Stroke, Other neurological, COPD, DU, DC, Hypothyroidism, Chronic kidney disease, Liver disease, Peptic ulcer, Aids, Lymphoma, Metastatic tumors, Solid tumor, Rheumatoid arthritis, Coagulopathy, Obesity, Anemia, Alcohol abuse, Drug abuse psychoses, Depression, Electrolyte and fluid disorder, Psychoses
Vital signs	Heart rate, Respiratory rate, Temperature, Systolic blood pressure, Diastolic blood pressure, Mean arterial pressure, Percutaneous oxygen saturation
Laboratory results	WBC, Hemoglobin, Platelet, Arterial PH, $PaCO_2$, PaO_2, Base excess, Na^+, Ca^+, K^+, Lactic acid, Cl^-, CVO_2, Blood glucose, Creatinine, BUN, Troponin, Total protein, BNP, CRP, GPT, GOT, TB, Albumin, Prothrombin time, APTT, INR
Ventilator information	DHMV, V_T, Minute volume, Mean airway pressure, PIP
Clinical intervention	Sedation day, Vasopressor
Clinical scores	RSBI, Sequential Organ Failure Assessment (SOFA), Glasgow Coma Scale (GCS)

COPD = chronic obstructive pulmonary disease, DU = Diabetes uncomplicated, DC = Diabetes complicated, , WBC = White blood cell, $PaCO_2$ = Arterial partial pressure of carbon dioxide, PaO_2 = Arterial partial pressure of oxygen, CVO_2 = Central vein oxygenated, BUN = Blood urea nitrogen, BNP = Brain natriuretic peptide, CRP = C-reactive protein, GPT = Glutamic-pyruvic transaminase, GOT = Glutamic oxalacetic transaminase, TB = Total bilirubin, APTT = Activated partial thromboplastin time, INR = International Normalized Ratio, DHMV = Duration hours of mechanical ventilation, PIP = Peak inspiration pressure.

discussion about important features and visualization of these features. Finally, conclusions and future works are described in Section V.

II. METHOD

A. DATA

We utilize the MIMIC-III clinical database, a freely accessible critical care data with 54379 admissions, for this study. The data includes patient demographic information, vital signs, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure and diagnostic codes, imaging reports, and so on [24]. To efficiently and deeply study the EF task, we selected 3636 patients from the initial admissions as follows. First, we excluded the patients who did not intubate and had unclear extubation states. Second, since our focus is on the adult ICU patients, the patients with age < 18 were removed. Third, admissions that the patient was dead before extubation were filtered out. Finally, 3636 unique adult patients undergoing mechanical ventilation and extubation were analyzed in our study and the numbers of extubation failure and extubation success are 624 and 3012 respectively.

According to previous studies, extubation failure was defined as the need for re-intubation within 48 hours after extubation [2]. Since direct information of whether the patient is intubated or not is not available in MIMIC-III, we infer the ventilation status (i.e., "intubated" or "not-intubated") through ventilator related parameters, such as respiratory support, ventilator mode, airway, breath rate, and oxygen delivery device [18]. In other words, if the patient has any of the parameters associated with using ventilator, the ventilation status of the patient is considered as intubated. An "extubation" event was defined as the time point where "intubated" status changed to 'not-intubated".

B. FEATURE PREPROCESSING

1) INITIAL FEATURES

Table 1 shows a preliminary set of features or variables that we used for this study. Note that these variables are selected according to the suggestions of our cooperated clinicians, and all of them are considered helpful by clinicians to predict EF. So we give a systematic and comprehensive analysis for the factors that may influence extubation outcomes. In total, we have 92 features, which contain demographic characteristics of the patient (e.g., age, gender and some past diseases), vital signs such as heart rate and respiratory rate, laboratory results for blood gas analysis, blood glucose and total protein, ventilator information including duration of mechanical ventilation, V_T and Minute volume, and clinical intervention such as sedation days and vasopressor. Also, the variables contain clinical scores of RSBI, SOFA and GCS. If a patient is extubated, we find and extract the above variables for the patient from the timeline before extubation.

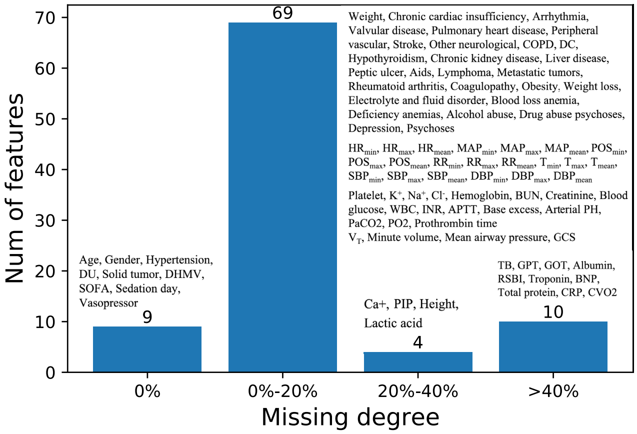


FIGURE 1. **Missing degree of initial features, where HR = Heart rate, MAP = Mean arterial pressure, POS = Percutaneous oxygen saturation, T = Temperature, SBP = Systolic blood pressure, and DBP = Diastolic blood pressure.**

2) MISSING VALUE PROCESSING

Fig. 1 demonstrates the missing degree of our initial features, and we divide the features into four groups: the feature values are not missing ("0%"), missing degree is less than 20% but greater than 0% ("0%-20%"), missing degree is in

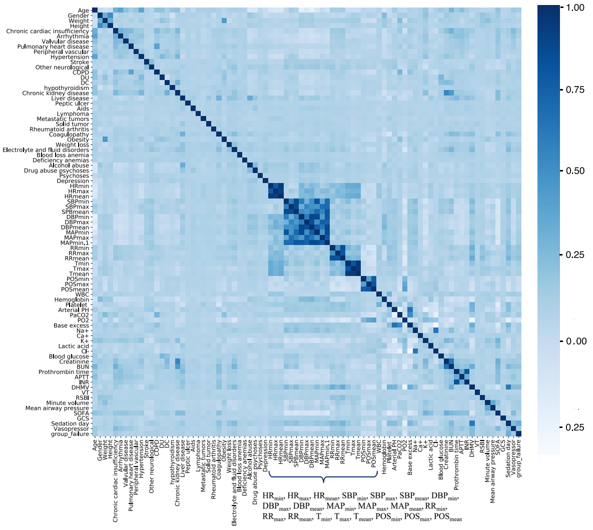


FIGURE 2. **Correlation analysis.**

20%-40% (“20%-40%”), and missing degree is greater than 40% (“>40%”). We remove features that their missing degree is greater than 40%. For features in group “20%-40%” (i.e., Ca^{+}, PIP, Height, and Lactic acid), we analyze the distributions of the values for each features, and the missing values for feature “Lactic acid” are imputed using the median of the “Lactic acid” values from other patients. The missing values for features “Ca^{+}”, “PIP” and “Height” are imputed similarly but using the mean values. Features in group “0%-20%” also use their respective means to fill the missing values.

3) CORRELATION ANALYSIS

To avoid different features having high correlation, we perform correlation analysis for the features after missing value processing. Pearson’s correlation coefficient is calculated for each feature against every other feature. The visualization result of correlation analysis is shown in Fig. 2. We find that features with high correlations ($R^2 > 0.75$) are the minimums, maximums and means for Hear rate, Respiratory rate, Mean arterial pressure, Percutaneous oxygen saturation, Temperature, Systolic blood pressure and Diastolic blood pressure, respectively. According to suggestions of clinicians, we choose the means for features HR, MAP, T, SBP and DBP, maximum for RR and minimum for POS. After the missing value processing and correlation analysis (removing 10 and 14 features respectively), we finally have 68 features for model input.

C. LIGHTGBM-BASED EF PREDICTION

1) LightGBM

LightGBM [25] is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks. In essence, LightGBM is an ensemble method that combines the predictions of multiple decision trees (by adding them together) to make the final prediction that generalizes well. Importantly, LightGBM trains the multiple tree models in an additive manner, with each new tree model being trained to predict the residuals (i.e., errors) of the prior models. Suppose we want construct a LightGBM model with T trees, and for a given dataset with n examples, the additive training process can be described as:

$$
\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
\tag{1}
$$

where $\hat{y}_i^{(t)}$ is the prediction of the i-th example at the t-th iteration and f_t is the learned function for the t-th decision tree. As the Eq. 1 illustrates, in each iteration, we keep the current model $\hat{y}_i$ and add a new function f (or the learned residuals) into the model. The fs of all iterations can be learned by minimizing the following objective:

$$
\mathcal{L}^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T} \Omega(f_t) \tag{2}
$$

The first term is the loss function measuring the difference between the prediction $y_i^{(t)}$ and the target y_i, and the second is the regularization term which penalize the complexity of the model.

Particularly, LightGBM is an implementation of gradient boosting decision tree (GBDT) [26]. When training each individual decision tree (f) and splitting the data, there are two exclusive strategies LightGBM employed: **gradient-based one-side sampling (GOSS)** [25] and **leaf-wise growth**. **GOSS** aims to tackle the computational complexity problem of conventional implementations of GBDT, which need to go through every feature of every data point when computing the information gain for all the possible splits. The crucial observation behind GOSS is that data instances with larger gradients play greater roles in information gain computation. Therefore, when estimating the best split, GOSS keeps data instances with large gradients and randomly samples data with small gradients. This strategy has been proved effective and work faster than conventional ones. **Leaf-wise growth** is an efficient strategy for growing trees. It finds the leaf with the largest splitting gain from all the current leaves each time, and then splits the leaf, and circulate this process. In other words, It will choose the leaf with max delta loss to grow. Compared with level-wise growth strategy, leaf-wise one can reduce more errors and obtain better accuracy under the same splitting times. The disadvantage of leaf-wise strategy is that it may grow trees deeply and lead to overfitting. Therefore, LightGBM adds a maximum depth limit on leaf-wise to ensure high efficiency while preventing overfitting. An illustration of leaf-wise and level-wise tree growth strategies is

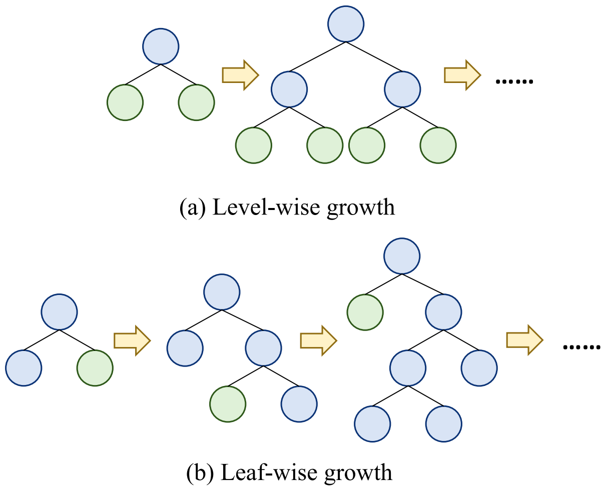


FIGURE 3. Strateges of tree growth.

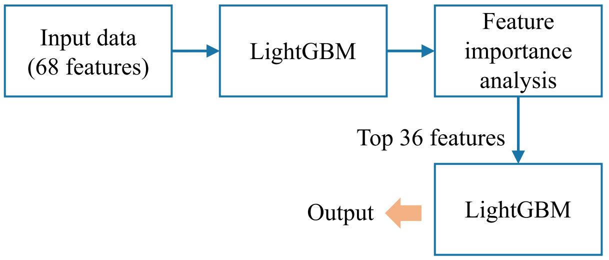


FIGURE 4. Extubation failure prediction process.

TABLE 2. Performance comparison. ACC = accuracy, SEN = sensitivity, and SPE = specificity.

Features	Models	ACC	SEN	SPE	AUC
68	LightGBM	0.8023	**0.7485**	**0.8327**	**0.8130**
	XGboost	0.7690	0.7468	0.7805	0.8114
	LR	**0.8137**	0.5158	0.8219	0.5285
	SVM	0.7828	0.5267	0.7731	0.6221
	ANN	0.7935	0.6247	0.8161	0.6836
36	LightGBM	**0.8020**	**0.7477**	**0.8394**	**0.8198**
	XGboost	0.7877	0.7454	0.8094	0.8168
	LR	0.7931	0.5346	0.7917	0.5425
	SVM	0.7734	0.5452	0.7834	0.6712
	ANN	0.8003	0.6865	0.7974	0.7358

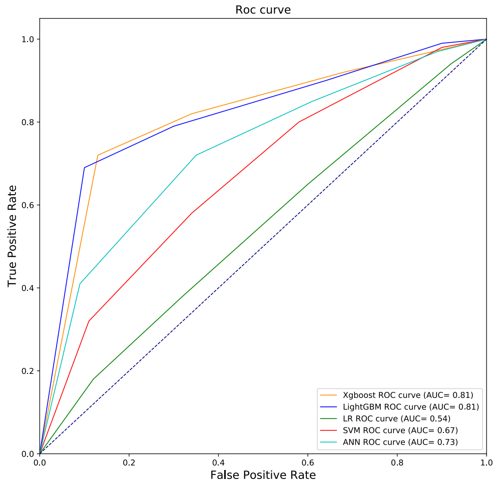


FIGURE 5. The result of feature importance analysis of LightGBM.

shown in Fig. 3. We discuss the impact of these two strategies on feature importance analysis in Section III-B.

2) EF PREDICTION

Due to the high performance and fast speed of LightGBM, in this paper, we explore the LightGBM method on EF prediction and analyze important factors that useful for EF prediction. Fig. 4 describes the whole process of our EF prediction has four steps:

(1) Before forming the final input data (with 68 features), we perform data processing including extracting eligible patients from the MIMIC-III clinical database, initial features extraction, missing value processing and correlation analysis.

(2) The LightGBM is applied on the input data. We set parameters "num_leaves =70" and "max_depth=6" to avoid over-fitting, especially the "max_depth" (avoid constructing trees too deep). The "min_data_in_leaf" parameter indicates the minimum number of samples per leaf node. It is an important parameter to deal with over-fitting of leaf-wise grown trees, and we set it to 30. In addition, the "learning_rate" is set to 0.02, and we use 5-fold cross-validation to compute the average prediction results.

(3) According to the result of step (2), we perform feature importance analysis, and find the top 36 features that are helpful for EF prediction. The feature importance ranking is based on the importance type "split" (in "feature_importance" function), which computes numbers of the times the feature is used in LightGBM to represent the importance of that feature.

(4) Data set with the analyzed top 36 features are inputted to a new LightGBM model, outputting the prediction for extubation failure. The final result is also obtained through 5-fold cross-validation.

D. EVALUATION METRICS

We evaluate the prediction performance of all the methods in terms of accuracy, sensitivity and specificity. The equations are described as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

where TP, TN, FP, FN refer to true positive, true negative, false positive, and false negatives, respectively. Additionally, receiver operating characteristic (ROC) curve as well as area under the ROC curve (AUC) are measured. The AUC value is computed by taking the integral of true positive rate with

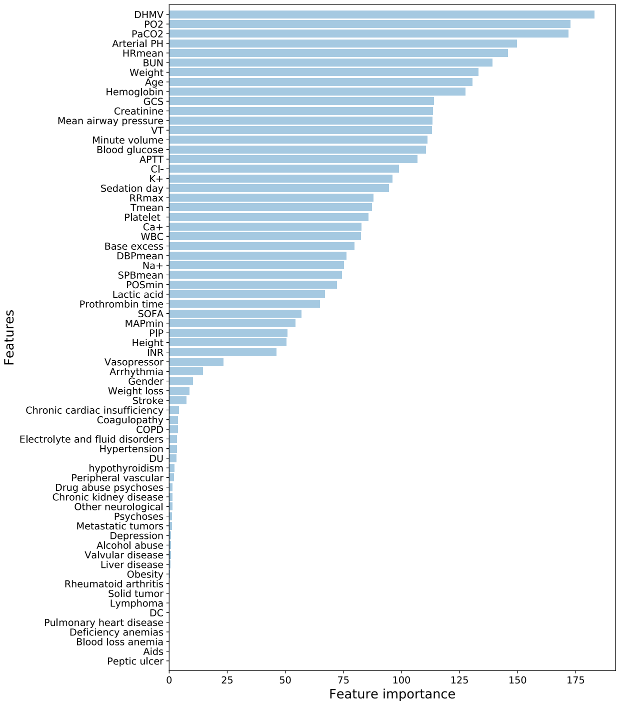


FIGURE 6. The feature importance ranking of LightGBM (with 68 features).

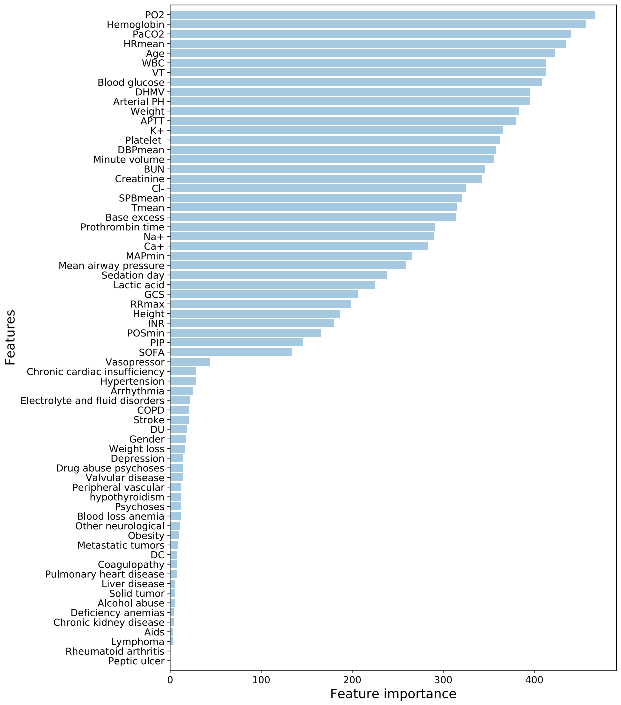


FIGURE 7. The feature importance ranking of XGboost (with 68 features).

respect to the false positive rate:

$$AUC = \int_0^1 R_{tp}(R_{fp})\delta R_{fp} \tag{6}$$

where the true positive rate R_{tp} is a function of the false positive rate R_{fp} along the curve.

III. RESULTS

A. PERFORMANCE OF LIGHTGBM AND FEATURE IMPORTANCE ANALYSIS

The performance of LightGBM and the result of feature importance analysis are shown in Table 2 and Fig. 6. Firstly, we apply the final input data with 68 features on LightGBM, which has an accuracy of 80.23%, sensitivity of 74.85%, specificity of 83.27% and the AUC value is 0.8130. Then, we perform feature importance analysis. As shown in Fig. 6, the three most important features are DHMV, PO_2 and $PaCO_2$, then the importance of features Arterial PH, HR_{min}, BUN, Weight, Age and Hemoglobin reduces but still maintains a certain importance. Further, we observe that almost all the five features related to ventilator information show high importance, especially the DHMV, Mean airway pressure, V_T and Minute volume. Features in laboratory results are also crucial for EF prediction, such as Hemoglobin, Blood glucose, APTT, K^+, and Cl^-. Since many features have little importance on the EF prediction, we extract the top 36 features for further analysis. As we can see, LightGBM with the top 36 features represents similar performance with that of 68 features, reducing slightly in accuracy and sensitivity but showing slight improvements in specificity and AUC value. This indicates that the removed 34 features are not such important and even not working in EF prediction.

B. PERFORMANCE COMPARISON WITH XGBOOST

XGboost [27] is also an implementation of GBDT. Comparing with LightGBM, the differences are that it adopts the **pre-sorting method** to compute the spitting information gain and the **level-wise method** (see Fig. 3(a)) to grow the tree. However, when computing the splitting gain, pre-sorting method needs to go through all the splitting points, consuming a lot of time and space. Further, level-wise tree construction method splits all the nodes in each layer equally. Some leaf nodes have a little splitting gain, which may not effect the results, but XGboost still splits, increasing the computational cost. The performance comparison between LightGBM and XGboost is shown in Table 2. For input 68 features, XGboost presents similar results to LightGBM on the sensitivity and AUC value, but decreases by 3.33% and 5.22% on the accuracy and specificity, respectively. Fig. 7 illustrates the feature importance analysis result of XGboost. As we can see, the important features for EF prediction is PO_2, Hemoglobin, $PaCO_2$, WBC and Blood glucose in Laboratory results, HR_{min} and DBP_{min} in vital signs, and Age and Weight in demographic information. Different from the result of LightGBM, the importance result of features in ventilator information of XGboost are not that important. Our clinicians indicate that the importance ranking of XGboost is not so accurate even it has the same top 36 features with LightGBM.

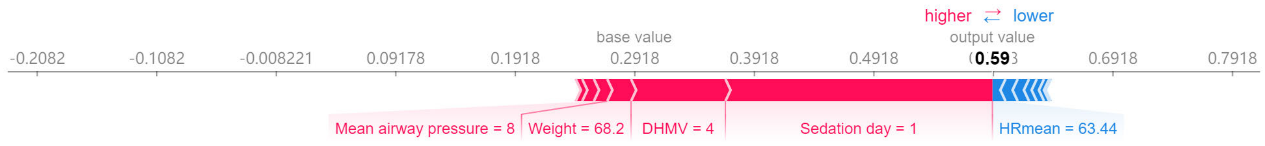


FIGURE 8. The SHAP value for a single data sample.

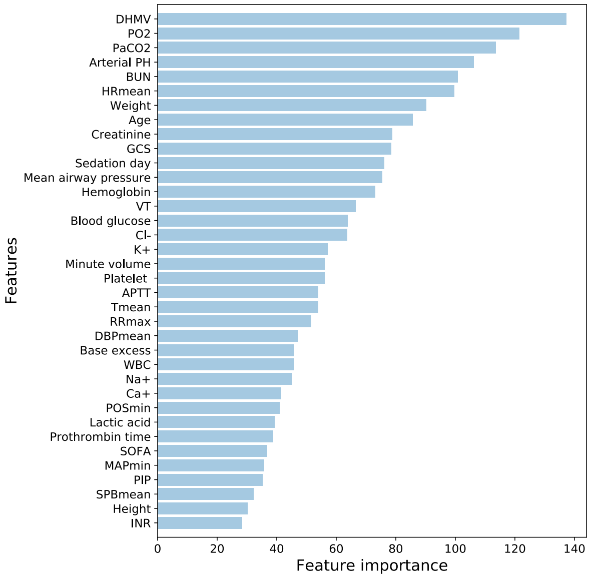


FIGURE 9. The feature importance ranking of LightGBM (with 36 features).

This is also the reason that we are not using XGboost for EF prediction.

C. PERFORMANCE COMPARISON WITH OTHER MACHINE LEARNING METHODS

Moreover, existing methods for EF prediction have utilized other machine learning methods, including LR [18], SVM [28] and ANN [17], [19]. We compare the EF prediction result of LightGBM with those of the aforementioned methods shown in Table 2 and Fig. 5, and we calculate the averaged accuracy, sensitivity, specificity and AUC value in the same setting as LightGBM (using 5-fold cross-validation). The LightGBM is generally superior to methods such as LR, SVM and ANN. All these methods show good results in accuracy and specificity, especially the LR with 68 features (81.37% accuracy and 82.19%) and ANN (80.23% accuracy) with 36 features, but they have pretty lower sensitivities and AUC values. The ROC curves in Fig. 5 also illustrate the superiority of LightGBM.

IV. DISCUSSION

After using LightGBM with the top 36 features to predict EF, we again perform feature importance analysis based on the trained LightGBM. The importance ranking is demonstrated in Fig. 9. Generally, comparing with LightGBM with 68 features, the importance orders of these features are not changing greatly, except for the feature Sedation day.

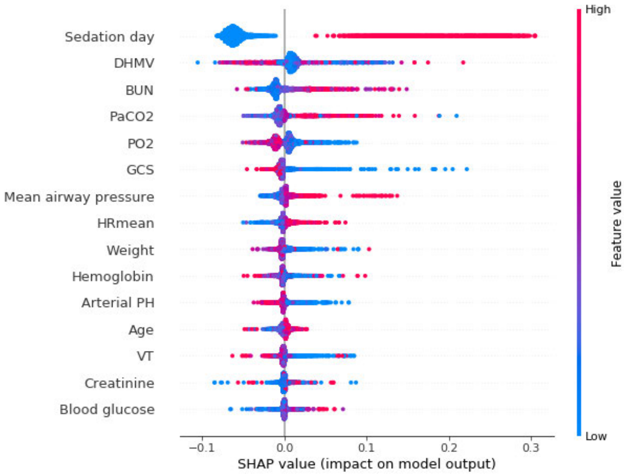


FIGURE 10. Overall analysis of features.

In this section, we go step further on the feature importance analysis. Feature importance is the contribution of each feature to improving the predictive ability of the entire model. It can intuitively reflect the importance of features and see which features have great influence on the final model, but it is impossible to judge how the relationship between the feature and the final prediction is. In Fig. 9, we can observe that the features of DHMV, PO_2 and $PaCO_2$ are the three most important factors affecting the prediction for EF. However, whether these features have positive or negative, or other more complex correlations with EF, it can not obtained from Fig. 9. SHAP (SHapley Additive exPlanations) [29], a interpretable tool for prediction output of machine learning models, can reflect the influence of the features in each data sample, and also shows the positive and negative effects. Therefore, we explore further on the correlation between features and EF using SHAP value and top 15 features in Fig. 9. The visualization results are shown in Fig. 10 and Fig. 8. Fig. 8 visualizes the SHAP values for a single data sample. The base value (0.2918) is the mean of the fitting values of targets in the training set. The blue indicates a negative contribution, and the red indicates a positive contribution. The below red texts are the values for these features, and the output value (0.58) is the prediction output of the sample. As shown in Fig. 8, the longest red bar is the Sedation day, and it increases the prediction value by 0.2 at least. The longest blue bar is the HR_{min}, and a small value of HR_{min} decreases the prediction result.

In Fig. 10, each row represents a feature with the SHAP value on the horizontal axis and each point represents a data sample. The redder the color means the larger the value of the feature itself, and the bluer the color means the smaller the value of the feature itself. The left features in Fig. 10 are re-sorted by the SHAP framework. From Fig. 10, we can intuitively observe that the Sedation day is a very important feature, and basically it is positively correlated with EF. Sedation day means the time of intubation patients using sedative drugs, which are crucial to maintain physiological stability and control pain levels of patients while intubated. The longer the sedation time also means the longer duration of mechanical ventilation, which makes extubation more difficult. The features of BUN, $PaCO_2$, Mean airway pressure, and HR_{min} are also positive to predict EF, and increasing the values of these features can increase the prediction of EF. The features of PO_2, GCS, Weight, Hemoglobin, Arterial PH, and V_T are negatively correlated with EF, and the smaller the values of these feature, the better the prediction of the model. Other features like DHMV, Creatinine and Blood glucose have both positive and negative correlations with EF, but DHMV prefers to negative correlation, Blood glucose prefers to positive correlation and Creatinine is neutral. For the feature Age, small feature values can reduce the prediction for EF, but large feature values can not only increase the prediction but also reduce it.

In addition, we also try the synthetic minority oversampling technique (SMOTE) [30] to enlarge the data of EF to overcome the imbalanced learning problem, but the final results with or without SMOTE are not obvious. So, we are not using SMOTE finally.

V. CONCLUSION

In this paper, we applied LightGBM to predict extubation failure using 3636 adult patient records in the MIMIC-III clinical database. We also performed a comprehensive feature importance analysis to identify useful features for EF prediction. When training the LightGBM, we firstly utilized the processed 68 features as input, and selected the top 36 features according to the result of feature importance analysis, and performed further study on these 36 feature. Experimental results demonstrated that using LightGBM is feasible for EF prediction and it outperformed other machine learning methods such as XGboost, LR, SVM and ANN. More importantly, we adopted the SHAP to explore the correlation between features and the prediction for EF. The feature importance and SHAP analysis illustrated that features of DHMV, PO_2 and $PaCO_2$ had great influence on the final model, and the Sedation day showed high positive correlation with EF, respectively. Of course, other features like Arterial PH, BUN, HR, Age and Weight were also important. In clinical practice, clinicians can concentrate more on those features to make more informed extubation decision. However, even though the LightGBM showed a better performance than other methods, the sensitivities (the recognition for EF) were still not very high. Further studies to improve the EF precition performance are needed.

REFERENCES

[1] J. F. McConville and J. P. Kress, "Weaning patients from the ventilator," *New England J. Med.*, vol. 367, no. 23, pp. 2233–2239, 2012.

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

[3] H. R. Hemant, J. Chacko, and M. K. Singh, "Weaning from mechanical ventilation-current evidence," *Indian J. Anaesth*, vol. 50, no. 6, pp. 435–438, 2006.

[4] P. Casaseca-de-la-Higuera, M. Martín-Fernández, and C. Alberola-López, "Weaning from mechanical ventilation: A retrospective analysis leading to a multimodal perspective," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 7, pp. 1330–1345, Jul. 2006.

[5] W. M. Coplin, D. J. Pierson, K. D. Cooley, D. W. Newell, and G. D. Rubenfeld, "Implications of extubation delay in brain-injured patients meeting standard weaning criteria," *Amer. J. Respiratory Crit. Care Med.*, vol. 161, no. 5, pp. 1530–1536, 2000.

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

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

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

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

[10] C. W. Seymour, A. Martinez, J. D. Christie, and B. D. Fuchs, "The outcome of extubation failure in a community hospital intensive care unit: A cohort study," *Crit. Care*, vol. 8, no. 5, pp. R322–R327, 2004.

[11] S. Epstein, "Decision to extubate," *Intensive Care Med.*, vol. 28, no. 5, pp. 535–546, 2002.

[12] B. Blackwood, F. Alderdice, K. Burns, C. Cardwell, G. Lavery, and P. O'Halloran, "Use of weaning protocols for reducing duration of mechanical ventilation in critically ill adult patients: Cochrane systematic review and meta-analysis," *BMJ*, vol. 342, p. c7237, Jan. 2011.

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

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

[15] B. Saugel, P. Rakette, A. Hapfelmeier, C. Schultheiss, V. Phillip, P. Thies, M. Treiber, H. Einwächter, A. von Werder, R. Pfab, F. Eyer, R. M. Schmid, and W. Huber, "Prediction of extubation failure in medical intensive care unit patients," *J. Crit. Care*, vol. 27, no. 6, pp. 571–577, 2012.

[16] J. A. Chaparro and B. F. Giraldo, "Power index of the inspiratory flow signal as a predictor of weaning in intensive care units," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 78–81.

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

[18] A. Mikhno and C. M. Ennett, "Prediction of extubation failure for neonates with respiratory distress syndrome using the MIMIC-II clinical database," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug./Sep. 2012, pp. 5094–5097.

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

[20] L. J. Kanbar, C. C. Onu, W. Shalish, K. A. Brown, G. M. Sant'Anna, D. Precup, and R. E. Kearney, "Undersampling and bagging of decision trees in the analysis of cardiorespiratory behavior for the prediction of extubation readiness in extremely preterm infants," in *Proc. IEEE 40th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 4940–4944.

[21] C. C. Onu, L. J. Kanbar, W. Shalish, K. A. Brown, G. M. Sant'Anna, R. E. Kearney, and D. Precup, "Predicting extubation readiness in extreme preterm infants based on patterns of breathing," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov./Dec. 2017, pp. 1–7.

[22] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks for Perception*. Amsterdam, The Netherlands: Elsevier, 1992, pp. 65–93.

[23] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Med.*, vol. 39, no. 5, p. 952, 2011.

[24] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, May 2016, Art. no. 160035.

[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.

[26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[28] M. Mueller, C. C. Wagner, R. Stanislaus, and J. S. Almeida, "Machine learning to predict extubation outcome in premature infants," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–6.

[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

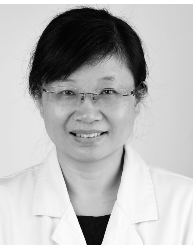
TINGTING CHEN received the B.S. degree in computer science from Southwest University, in 2017. She is currently pursuing the Ph.D. degree with the College of Computer Science, Zhejiang University. Her current research interests include machine learning, deep learning, computer vision, and medical intelligence.

JUN XU received the M.M. degree from Zhejiang University, in 2014, where he is currently with the Intensive Care Unit, The First Affiliated Hospital, College of Medicine. As an attending Physician, he has published two research articles in medical journals. His current research interests include data extraction and data analysis.

HAOCHAO YING received the B.S. degree in computer science from the Zhejiang University of Technology, in 2014. He is currently pursuing the Ph.D. degree with the College of Computer Science, Zhejiang University. His current research interests include recommender system and data mining, especially on digital health care and location-based social network.

XIAOJUN CHEN received the M.S. degree in applied informatics from Liverpool University, in 2017. He is currently with the Real Doctor AI Research Centre, Zhejiang University. His current research interests include data mining, machine learning, and deep learning.

RUIWEI FENG received the B.S. degree from North China Electric Power University, in 2018. She is currently pursuing the Ph.D. degree with the College of Computer Science, Zhejiang University. Her current research interests include machine learning, deep learning, computer vision, and medical intelligence.



XUELING FANG received the M.D. degree from Zhejiang University. She was with the Intensive Care Unit, The First Affiliated Hospital, College of Medicine, Zhejiang University, for more than 20 years. As a Chief Physician and a Deputy Director of the department, she has extensive clinical experience in treating critically ill patients. She has published several research articles in international journals such as *CHEST*. Her current research interests include sepsis, immunology, stem cell, and clinical big data.

HONGHAO GAO received the Ph.D. degree in computer science and started his academic career with Shanghai University, in 2012. He is currently a Distinguished Professor with the Key Laboratory of Complex Systems Modeling and Simulation, Ministry of Education, China. His current research interests include service computing, model checking-based software verification, and sensors data application. He is an IET Fellow, BCS Fellow, EAI Fellow, CCF Senior Member, and CAAI Senior Member.

JIAN WU received the Ph.D. degree in computer science and technology from Zhejiang University, in 1998. He is currently the Director of the Real Doctor AI Research Centre and the Vice-President of the National Research Institute of Big Data of Health and Medical Sciences, Zhejiang University. His current research interests include medical artificial intelligence, service computing, and data mining. He is a CFF member, CCF TCSC member, CCF TCAPP member, and member of the 151 Talent Project of Zhejiang Province.