Business Administration Master's Degree Thesis

Using retrospective analysis of the MIMIC-III database and comparison of machine learning algorithms to predict for weaning failure of mechanical ventilation

February 2022

Seoul School of Integrated Science and Technology

Sujung Lee

Business Administration Master's Degree Thesis

Using retrospective analysis of the MIMIC-III database and comparison of machine learning algorithms to predict for weaning failure of mechanical ventilation

February 2022

Seoul School of Integrated Science and Technology

Sujung Lee

Using retrospective analysis of the MIMIC-III database and comparison of machine learning algorithms to predict for weaning failure of mechanical ventilation

Thesis Supervisor Joong ho, Chang

This paper is submitted as the Business Administration Master's Dissertation

February 2022

Seoul School of Integrated Science and Technology
Sujung Lee

This approves Sujung Lee's Master's Dissertation

January 2022

Committee Chair Bo Young, Kim (signature)

Committee Dal Ju, Mun (signature)

Committee Joong ho, Chang (signature)

Abstract

Mechanical ventilation is a common life support system used in the intensive care unit (ICU). 48.8% of ICU patients were received invasive mechanical ventilation. Weaning is a process that is completely independent of mechanical ventilation. Extubation is the last step in weaning, and extubation failure is defined as a case in which the patient is not completely independent from mechanical ventilation. Weaning failure is an important issue in mechanical ventilation. Reintubation after extubation due to weaning failure has reached 6%-47%, which puts a physical burden on the patient. To assess the appropriate extubation timepoint, several clinical weaning indices are used, such as RSBI, Pimax, P0.1, and P0.1/Pimax. However, no perfect index is available that can be used to determine weaning success. This research examines various machine learning models and aim to compare a model that predicts extubation failure more accurately.

The Medical Information Mart for Intensive Care (MIMIC-III) database was used as the data resource. MIMIC-III is a single-center database covering 38,597 distinct adult patients admitted to the ICU in the Beth Israel Deaconess Medical Center in Boston from 2001 to 2012. I selected subjects who met eligibility (n=2,094) and

extracted data. There were multiple hospitalizations per patient, resulting in a total of 3,942 cases collected. All missing values were removed, and the data used for analysis were 747cases, which were randomly divided into training (80%) and testing (20%) datasets. Datasets are fitted in Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost and Light GBM Algorithm. Find best hyperparameter using 5-fold Gridsearch cross validation on each algorithm. And comparing the model performance results. The AUROC for extubation failure was 0.655, 0.900, 0.620, 0.850, 0.970, 0.966, 0.969. Feature importance of XGBoost and Light GBM, which had excellent results, was analyzed. The top 5 of features of each model were Ve(minute ventilation), GCS(Glasgow coma scale), Height, Vt(tidal volume) and OASIS(Outcome and Assessment Information Set) for XGBoost, and Ve, OASIS, HR(heart rate), SpO2(Saturation of percutaneous oxygen) and MBP(Mean Blood Pressure) for Light GBM. Ve(minute ventilation) had the highest impact power in XGBoost and Light GBM models.

Keywords: mechanical ventilation, weaning failure, machine learning

초록

기계 환기는 중환자실(ICU)에서 사용되는 일반적인 생명 유지 시스템입니다. ICU 환자의 48.8%가 침습적 기계환기를 받습니다. 이탈은 기계환기와 완전히 독립하는 과정입니다. 발관은 이탈의 마지막 단계이며 발관 실패는 환자가 기계적 환기로부터 완전히 독립되지 않은 경우로 정의합니다. 이탈 실패는 기계환기에서 중요한 문제입니다. 이탈 실패로 발관 후 재삽관은 6~47%에 이르며 환자에게 육체적인 부담을 줍니다. 적절한 발관 시점을 평가하기 위해 RSBI, Pimax, P0.1, P0.1/Pimax 와 같은 여러 임상적 이탈 지표가 사용됩니다. 그러나 이탈 성공을 결정하는 데 사용할 수 있는 완벽한 지표는 없습니다. 본 연구는 다양한 머신러닝 모델을 살펴보고 발관 실패를 보다 정확하게 예측하는 모델을 비교하는 것을 목적으로 합니다.

데이터는 MIMIC-III (Medical Information Mart for Intensive Care) 데이터베이스를 사용하였습니다. MIMIC-III 는 2001 년부터 2012 년까지 보스턴에 있는 Beth Israel Deaconess Medical Center 의 중환자실에 입원한 38,597 명의 개별 성인 환자를 다루는 단일 센터 데이터베이스 입니다. 적격성을 충족하는 대상자를 선택하고 데이터를 추출했습니다(n=2,094). 환자 1 인당 여러 번 입원하여 총 3,942 건의 증례가 수집되었습니다. 모든 결측값을 제거하고 분석에 사용된 데이터는 747 건으로 훈련 데이터 세트(80%)와 테스트 데이터

세트(20%) 로 무작위로 분리했습니다. 데이터세트는 Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost, Light GBM 알고리즘에 적합 했습니다. 각 알고리즘에 대해 5fold-Gridsearch 교차 검증을 사용하여 최상의 하이퍼파라미터를 찾았습니다. 발관 실패에 대한 AUROC는 0.655, 0.900, 0.620, 0.850, 0.970, 0.966, 0.969 였습니다. 우수한 결과를 보인 XGBoost 및 Light GBM 모델의 특성 중요도를 분석했습니다. 각 모델의 상위 5 개 특성은 XGBoost 의 경우 Ve(분당환기량), GCS(글라스고우 혼수 척도), 신장, Vt(일회호흡량), OASIS(Outcome and Assessment Information Set)였으며 Light GBM 에서는 Ve, OASIS, HR(심박수), SpO2(산소포화도), MBP(평균동맥압) 이었습니다. Ve는 XG Boost 와 Light GBM 모델에서 가장 높은 영향력을 보였습니다.

키워드 : 기계 환기, 이탈 실패, 기계 학습

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Research Background and Objective

Mechanical ventilation is a common life support system used in the intensive care unit (ICU) and 48.8% of patients admitted to the ICU receive invasive mechanical ventilation (Metnitz, Metnitz, Moreno, & al., 2009). Mechanical ventilation can provide patients with respiratory failure with appropriate oxygenation and ventilation and give clinicians more time to treat the underlying disease (Wonsch.H., 2013). Disability of gas exchange due to lung disease, decreased consciousness due to cranial nervous system abnormalities, decreased function of the respiratory center due to cranial nervous system abnormalities, and hemodynamic instability due to cardiovascular disease are representative causes of mechanical ventilation (M.Tobin. & C.Manthous., 2017). Mechanical ventilation hinders normal blood flow circulation due to increased thoracic internal pressure, causes airway damage, airway pressure-related lung damage, pneumonia, and respiratory muscle atrophy, and prolonged mechanical ventilation leads to an extension of treatment period, mortality, and morbidity. Weaning is a concept that encompasses the entire process in which respiratory failure is corrected and the patient is freed from endotracheal tube,

mechanical assistance, and all related terminal care. After the acute phase has passed and the patient's respiratory status is stabilized, the clinician performs the Spontaneous Breathing Test (SBT) and finally proceeds to extubation by referring to various indicators. The process from SBT to extubation can be viewed as a departure from mechanical ventilation (Kyu-Hyouck, 2012).

Extubation is the last step in weaning, and extubation failure is defined as a case in which the patient is not completely independent from mechanical ventilation, and weaning failure is an important issue in mechanical ventilation. Premature weaning of mechanical ventilation can increase the burden on the respiratory and cardiovascular systems. Delayed weaning of mechanical ventilation may cause diaphragmatic atrophy and weakness (MJ., 2001), (Mahmood S, et al., 2014). Therefore, there is a need for a method that can reduce the mechanical ventilation period and respiratory damage caused by performing weaning of mechanical ventilation at an appropriate time.

To assess the appropriate extubation timepoint, several clinical weaning indices are used, such as the rapid shallow breathing index (RSBI), maximal inspiratory pressure(Pimax), airway occlusion pressure in the first 100ms(P0.1) and P0.1/Pimax, the most commonly used weaning index is the RSBI (S., et al., 2015).

However, no perfect index is available that can be used to determine weaning success. Reintubation due to extubation failure of mechanical ventilation is not uncommon (Mahmood S, et al., 2014).

Extubation failure is defined as the inability to sustain spontaneous breathing after removal of the endotracheal tube and the need for re-intubation within 48 hours of extubation. (Kuilkarni AP & Agarwal., 2008), (Mahmood S, et al., 2014). In several studies, reintubation rate is 6% to 47% (Epstein, Ciubotaru, & Wong, 1997), (RL. & RP., 1996), (Esteban, et al., 1995), (Brown CV, et al., 2011). The prognosis of patients who have experienced extubation failure is poor, and it is known that the in-hospital mortality rate is more than 30-40%. (Epstein SK & RL., 1998). Early prediction of extubation failure will substantially minimize the need for extended ventilators, long stays in the ICU, morbidity, mortality and financial burden on the health care system (Epstein SK & RL., 1998), (Seymour CW, Martinez A, Christie JD, & BD., 2004).

Therefore, when deciding on extubation, it is necessary to consider various factors along with the process of reducing dependence on mechanical ventilation and more accurate indicators to improve the current failure rate should be developed. This research examines various machine learning models and aims to compare a model that predicts extubation failure more accurately.

1.2 Research Design

This research is retrospective research. I screened the information of patients who performed invasive mechanical ventilation which was searched from MIMIC-III database with intubation item code and extubation item code. After that, I collected the eligible subject information according to the selection criteria which were aged 18 to less than 70 years old adult, the age calculated based on the date of hospital admission. According to the above criteria, subjects were recruited from the entire MIMIC-III database (n=29,547), and among them, those who maintained mechanical ventilation for 24 hours or more were extracted (n=2,094). Among the 2,094 subjects, there was a case where the subject experienced more than one invasive mechanical ventilation event due to more than one hospitalization. Thus, the number of mechanical ventilation cases of the extracted subjects(n=2,094) was 3,942 cases. These cases were divided into the extubation failure group and the extubation non-failure group. To divide the group, extubation failure was defined as reintubation within 48 hours.

Intubation, Extubation, Reintubation events of the above cases, Gender, Age, Height, SAPS-II(Simplified Acute Physiology Score-

II), OASIS(The Outcome and Assessment Information Set), GSC(Glasgow Coma Scale), SpO2(Saturation of percutaneous oxygen), FiO2(Fraction of inspired oxygen), MBP(Mean Blood Pressure), Vt(Tidal Volume), Ve(Minute Ventilation), HR(Heart Rate), RR(Respiratory Rate), Pimax(Maximal Inspiratory Pressure), PaCO2(Carbon dioxide pressure in arterial blood) and COPD(Chronic obstructive pulmonary disease) were collected. Outliers were removed from the collected data, and the removal criteria were deleted according to https://github.com/MIT-LCP/mimic-code.

Learning was performed with 747 cases data by removing missing values. To preserve the meaning of the original data, datasets were not imputing the missing data.

The best model would select by training and validation with Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost and Light GBM algorithms. Dataset divided randomly into training set (80%) and testing set (20%). The flow chart of the research designs is shown in Figure 1.

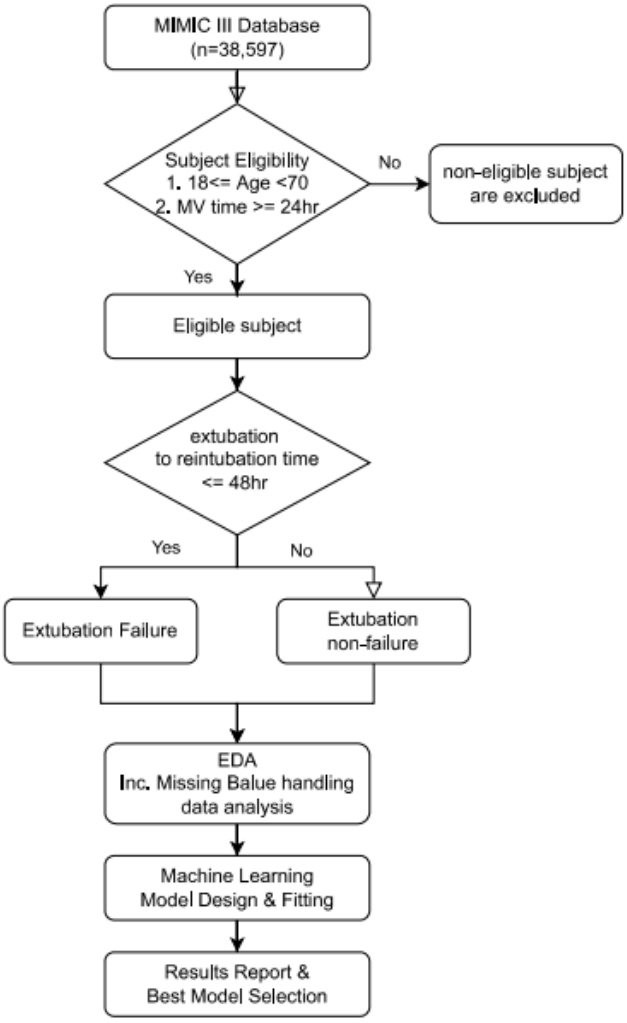


Figure 1 Flowchart of the research design

Chapter 2. Theoretical Background

2.1 Review of Weaning Failure Previous Studies

When searched PubMed and google scholar for studies published in

English from Jan 1, 2010, to Dec 30, 2021, using the search terms “mechanical ventilat wean machine leaning”. Chung WC’s research was the only article that directly predicts weaning using machine learning.

This research has limitation that only 169 patient ‘s data using for learning and Chung suggest only the neural network model which trained sex, height, oxygen saturation, Glasgow Coma Scale, Acute Physiology and Chronic Health Evaluation II score, pulmonary disease history, and respiratory parameters of the first, 30th, 60th, and 90th minute. (Chung WC, et al.) Using respiratory parameters of various time points could difficult to apply to the real world. Because each institution may have different facilities, systems, patient treatment protocols, and other resources, a predictive model that could be accessed more easily and learned using general variables is needed. Also, various machine learning methods, as well as neural networks, have the potential to suggest alternatives to create an optimal classifier. Chung’s research reported only neural networks, there is insufficient evidence to suggest that it is an optimal model for predicting extubation failure/success.

At Zhu’s research, the models were trained by features of the first day of admission using KNN, Logistic regression, bagging, decision tree, random forest, Extreme Gradient Boosting (XGBoost) and

Neural Network (feed-forward network). XGBoost was the best. various machine learning models were predicted the mortality of subjects who experienced mechanical ventilation, but Zhu's research was not studied about success of weaning or weaning failure (Zhu Yibing, et al., 2021).

Therefore, this research designed more subjects were recruited to increase power, selected minimized variables to expand applicability. The subject's indications were not disease-specific, and to conduct comparative analysis of various machine learning methods. Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGboost and Light GBM were selected to compare the best classification model.

2.2 Mechanical ventilation and Weaning

2.2.1 Mechanical ventilation

Mechanical ventilation is a form of life support. A mechanical ventilator is a machine that takes over the work of breathing when a person is not able to breathe enough on their own. There are many reasons why a patient may need a ventilator, but low oxygen levels or severe shortness of breath from an infection such as pneumonia are the most common reason. Ventilators are used to deliver high

concentrations of oxygen into the lungs, to help get rid of carbon dioxide, to decrease the amount of energy a patient uses on breathing so their body can concentrate on fighting infection or recovering, to breathe for a person who is not breathing because of injury to the nervous system, like the brain or spinal cord, or who has very weak muscles, to breathe for a patient who is unconscious because of a severe infection, buildup of toxins, or drug overdose.

When a person needs to be a ventilator, a healthcare provider will insert an endotracheal tube (ET tube) through the patient's nose or mouth and into their windpipe(trachea). This tube is then connected to the ventilator. The endotracheal tube and ventilator do a variety of jobs. The ventilator pushes a mixture of air and oxygen into the patient's lungs to get oxygen into the body.

The mechanical ventilation had several risks such as pneumonia, pneumothorax and other lung damages. (M.Tobin. & C.Manthous., 2017)

2.2.2 Weaning from Mechanical ventilation

Weaning from mechanical ventilation is an essential and universal element in the care of critically ill intubated patients receiving mechanical ventilation. Weaning covers the entire process of liberating the patient from mechanical support and from the

endotracheal tube, including relevant aspects of terminal care. There is uncertainty about the best methods for conducting this process, which will generally require the cooperation of the patient during the phase of recovery from critical illness. This makes weaning an important clinical issue for patients and clinicians.

Weaning is a series of stages in the process of care, from intubation and initiation of mechanical ventilation through the initiation of the weaning effort to the ultimate liberation from mechanical ventilation and successful extubation which has six stages as follows. Stage 1 was treatment of acute respiratory failure (ARF), stage 2 is suspicion that weaning may be possible, stage 3 is assessment of readiness to wean, stage 4 is Spontaneous breathing trial (SBT), stage 5 is extubation and possibly stage 6 is reintubation.

There is much evidence that weaning tends to be delayed, exposing the patient to unnecessary discomfort and increased risk of complications, and increasing the cost of care. Demonstrated that mortality increases with increasing duration of mechanical ventilation, in part because of complications of prolonged mechanical ventilation, especially ventilator–associated pneumonia and airway trauma. Moreover, mechanical ventilation costs ~US$2000 per day Subjects receiving prolonged mechanical ventilation account for 6% of all ventilated patients but consume

37% of intensive care unit (ICU) resources.

Increase in the extubation delay between readiness day and effective extubation significantly increase mortality. Mortality was 12% if there was no delay in extubation and 27% when extubation was delayed.

Thus, criteria for readiness to begin weaning should be systematically evaluated each day to allow prompt initiation of weaning as soon as the patient is ready this will shorten the weaning process and minimize time on mechanical ventilation. This is also an independent predictor of successful extubation and survival.

In most studies, weaning failure is defined as either the failure of SBT or the need for reintubation within 48h following extubation. Failure of SBT is defined by: one objective indices of failure, such as tachypnea, tachycardia, hypertension, hypotension, hypoxemia or acidosis, arrhythmia and subjective indices, such as agitation or distress, depressed mental status, diaphoresis and evidence of increasing effort.

Failure of extubation is associated with high mortality rate, either by selecting for high-risk patients or by inducing deleterious effects such as aspiration, atelectasis and pneumonia. Interestingly, mortality is not especially increased when failure of extubation is related to upper airway obstruction (one out of nine patients, 11%)

but is markedly increased in the other case(19 out of 52 patients : 36%) (J-M. Boles, 2007).

2.3 Machine Learning Classifiers

In this research, I compared and analyzed Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost and Light GBM models. The theoretical background of each machine learning model are as follows.

2.2.1 Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. There are two types of logistic regression, one is binary, the other is multi-linear functions fails Class. Classification models that directly specify class labels without calculating class conditional probabilities are called discriminative models. Logistic regression is a probabilistic discriminant model. It directly estimates the odds ratio of the data instance x using the attribute values. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

$$z = w^T x + b$$

Since logistic regression analysis has different weights for all properties, it is possible to understand the relationship between properties and class labels by analyzing the learned logistic regression parameters. Also, because logistic regression does not include computational density and distance in the attribute space, it can work more robustly in higher-dimensional settings than distance-based methods such as nearest-neighbor classifiers. However, the objective function of logistic regression does not include terms related to the complexity of the model. Therefore, logistic regression does not provide the same method as support vector machines. Nevertheless, variants of logistic regression can be easily developed to account for model complexity by including appropriate terms in the objective function along with the cross-entropy function. (Tan, 2019)

2.2.2 KNN

K-nearest neighbors (KNN) is a supervised machine learning algorithm that can be used to solve both classification and regression tasks. K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new

case/data and available cases and puts the new case into the category that is most like the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. KNN is a non−parametric algorithm, which means it does not make any assumption on underlying data.

K−nearest neighbor (KNN) classification, finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood.

$$\text{Majority Voting: } \acute{y} = \underset{v}{argmax} \sum_{(x_i y_{i)} \in D_z} I(v = y_i)$$

There are three key elements of this approach: a set of labeled objects such as a set of stored records, a distance or similarity metric to compute the distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabeled object, the distance of this object to the labeled objects is computed, its k−nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the

object.

(Wu, et al., 2008)

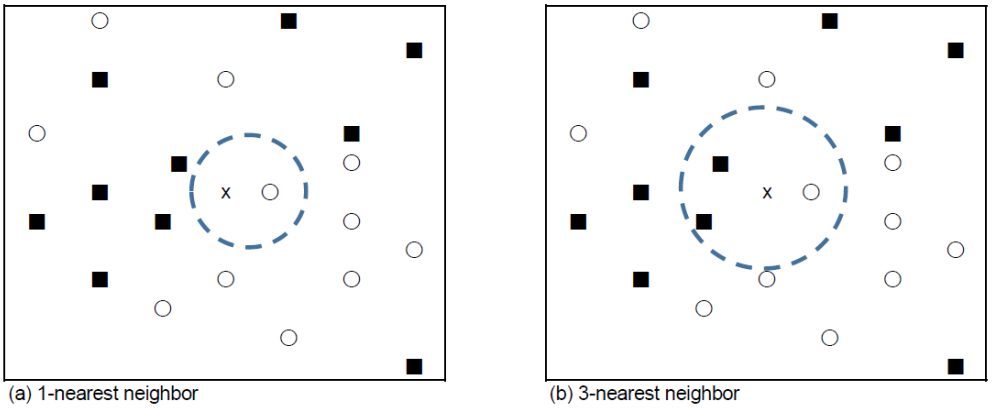


Figure 2 k nearest neighbor instance

2.2.3 SVM

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems. SVM algorithm creates a line or a hyperplane which separates the data into classes. At first approximation what SVMs do is to find a separating line (or hyperplane) between data of two classes. SVM is an algorithm that takes the data as an input and outputs a line that separates those classes if possible.

In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the

"best" classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane $f(x)$ that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance x_n can be classified by simply testing the sign of the function $f(x_n); x_n$ belongs to the positive class if $f(x_n) > 0$.

Because there are many such linear hyperplanes, what SVM additionally guarantee is that the best such function is found by maximizing the margin between the two classes. Intuitively, the margin is defined as the amount of space, or separation between the two classes as defined by the hyperplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an infinite number of hyperplanes, only a few qualify as the solution to SVM.

The reason why SVM insists on finding the maximum margin hyperplanes is that it offers the best generalization ability. It allows not only the best classification performance (e.g., accuracy) on the training data, but also leaves much room for the correct classification of the future data. To ensure that the maximum margin

hyperplanes are actually found, an SVM classifier attempts to maximize the following function with respect to $\vec{w}$ and b:

$$L_p = \frac{1}{2} \ ||\vec{w}|| - \ \sum_{i=1}^{t} \alpha_i y_i (\vec{w} \cdot \vec{x}_i + b) + \sum_{i=1}^{t} \alpha_i$$

(Wu, et al., 2008).

2.2.4 Decision Tree

Decision Tree is one of the predictive modeling algorithm. It uses a decision tree to go from observations about an item to conclusions about the item' s target value. Tree models where the target variable can take a discrete set of values are called classification trees, and Decision trees where the target variable can take continuous values are called regression trees. (Wu, et al., 2008)

Decision tree has a hierarchical structure consisting of nodes and directed edges. The tree has three types of nodes. A root node that has no incoming edges and zero or more outgoing edges. Each of internal nodes has exactly one incoming edge and two or more outgoing edges. Each leaf or terminal node has exactly one incoming edge and no outgoing edges. In a decision tree, each leaf node is assigned a class label. The non-terminal nodes include the root and other internal nodes.

Classifying a test record is straightforward once a decision tree has

been constructed. Starting from the root node, we apply the test condition to the record and follow the appropriate branch based on the outcome of the test.

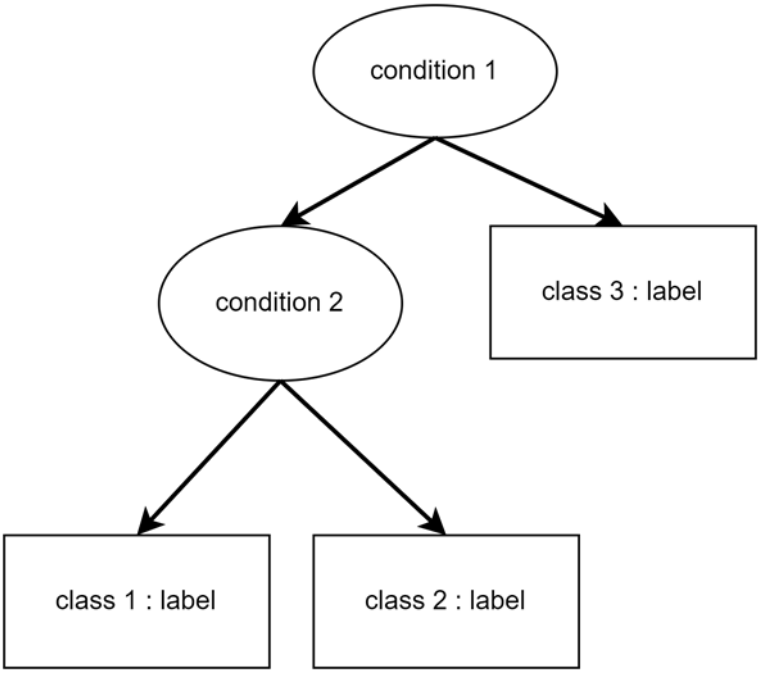


Figure 3 Classifying of decision tree

Decision tree induction is a nonparametric approach for building classification models and provides an expressive representation for learning discrete-valued functions. However, they do not generalize well to certain types of Boolean problems. Decision tree algorithms are quite robust to the presence of noise. Feature selection techniques can help to improve the accuracy of decision trees by eliminating the irrelevant attributes during preprocessing.

(Tan, 2019)

2.2.5 Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Random forest is a class of ensemble methods specifically designed for decision tree classifiers. It combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors, as shown in Figure 4.

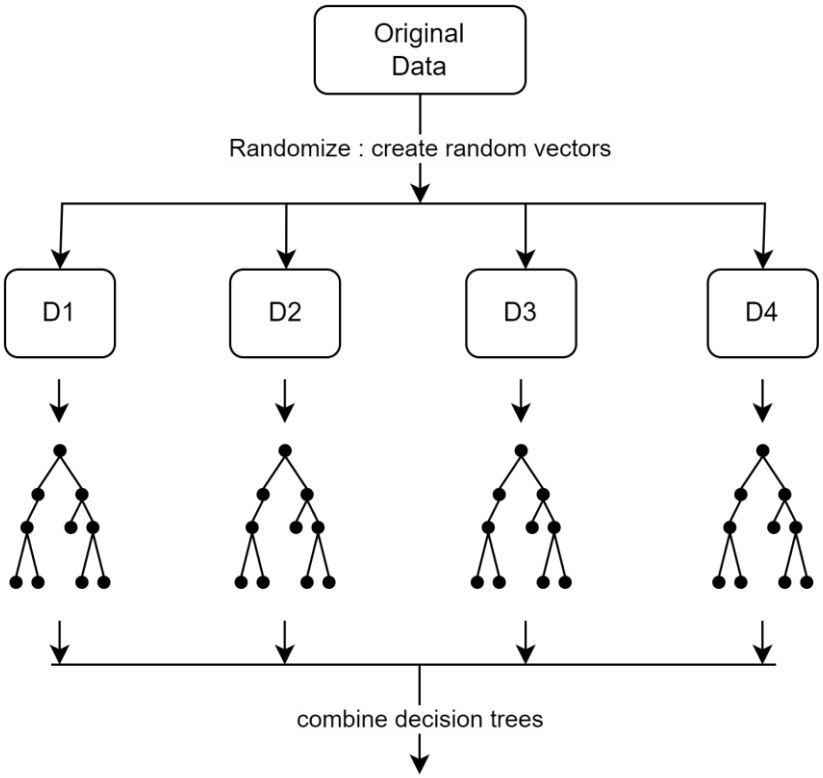


Figure 4 Random forests

The random vectors are generated from a fixed probability distribution, unlike the adaptive approach used in Ada boost, where the probability distribution is varied to focus on examples that are hard to classify. Bagging using decision trees is a special case of random forests, where randomness is injected into the model-building process by randomly choosing N samples, with replacement, from the original training set. Bagging also uses the

same uniform probability distribution to generate its bootstrapped samples throughout the entire model-building process. It was theoretically proven that the upper bound for generalization error of random forests converges to the following expression when the number of trees is sufficiently large.

Random forest improves generalization performance by constructing an ensemble of building decision trees. Random forest is based on the idea of bagging to use different bootstrap samples of training data to learn decision trees. However, the main feature of random forests that distinguishes it from bagging is that the best splitting criterion is chosen from a small set of randomly selected properties at all internal nodes of the tree. In this way, the random forest builds an ensemble of decision trees from training instance manipulation (using bootstrap samples like bagging) and input properties (using a different subset of properties at every internal node).

A bootstrap sample D_i of the training set is built by randomly sampling n instances (randomly replaced) from the dataset. We use D_i to train a decision tree T_i. We randomly sample a set of p attributes from all internal nodes of D_i and select the attribute that represents the greatest reduction in impurity measurements for isolation from this subset. Repeat this procedure until all leaves are

clean. When the ensemble of decision trees is constructed, the average prediction (representative vote) for the test instances is used as the final prediction of the random forest. The decision tree contained in the random forest is an unorganized tree, which can grow to the maximum possible size until all the leaves are pure.

Therefore, the basic classification criterion of the random forest is an unstable classifier with low bias but high variance due to its large size. Random forests can aggregate predictions for ensembles of strongly and unrelated decision trees, reducing the variance of the tree without negatively impacting low bias. This makes the random forest overfitted (Tan, 2019).

2.2.6 XGBoost

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now.

XGBoost is a scalable machine learning system for tree boosting. The system is available as an open-source package. XGBoost has

scalability in all scenarios. The system runs more than ten times faster than existing popular solutions on a single machine and scales to billions of examples in distributed or memory-limited settings.

The scalability of XGBoost is due to several important systems and algorithmic optimizations. One of those that is a novel tree learning algorithm is for handling sparse data and a theoretically justified weighted quantile sketch procedure enables handling instance weights in approximate tree learning.

In gradient tree boosting algorithms, the derivation follows from the same idea in existing literature in gradient boosting. Specifically, the second-order method is originated. We make minor improvements in the regularized objective, which were found helpful in practice. Regularized Learning Objective For a given data set with n examples and m features $\mathrm{D} = \{(x_i, y_i)\}(|\mathrm{D}| = \mathrm{n}, x_i \in R_m, y_i \in \mathrm{R})$, a tree ensemble model uses K additive functions to predict the output.

$$\widehat{y_i} = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F,$$

Gradient tree boosting

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t(X_i)\right) + \Omega(f_t)$$

This equation can be used as a scoring function to measure the quality of a tree structure q. This score is like the impurity score for evaluating

decision trees, except that it is derived for a wider range of objective functions

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma \mathrm{T}.$$

XGBoost proposed a novel sparsity aware algorithm for handling sparse data and a theoretically justified weighted quantile sketch for approximate learning. XGBoost is able to solve real-world scale problems using a minimal amount of resources (Chen & Guestrin, 2016).

2.2.7 Light GBM

Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm, used for ranking, classification and many other machine learning tasks.

Since it is based on decision tree algorithms, it splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. So, when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy which can rarely be achieved by any of the existing boosting algorithms.

GBDT (Gradient Boosting Decision Tree) is a popular machine

learning algorithm and has quite a few effective implementations such as XGBoost and pGBRT. However, the efficiency and scalability are still unsatisfactory when the feature dimension is high and data size is large. A major reason is that for each feature, they need to scan all the data instances to estimate the information gain of all possible split points, which is very time-consuming. To tackle this problem, Light GBM uses two technics. One is GOSS (Gradient-based One-side sampling) and the other is EFB(Exclusive Feature Bundling) GOSS can obtain quite accurate estimation of the information gain with a much smaller data size. EFB, which can bundle mutually exclusive features is NP-hard, but a greedy algorithm can achieve quite good approximation ratio. Therefore, the number of features can be effectively reduced without significantly compromising the accuracy of split-point determination. New GBDT implementation using GPSS and EFB is called Light GBM.

In Light GBM, GBDT uses decision trees to learn a function from the input space X^S to the gradient space $\mathbf{G}$. Suppose that we have a training set with $\mathbf{n}$ instances $\{x_1, \cdots, x_n\}$, where each x_i is a vector with dimension $\mathbf{s}$ in space X^S. In each iteration of gradient boosting, the negative gradients of the loss function with respect to the output of the model are denoted as $\{g_1, \cdots, g_n\}$. The decision tree model

splits each node at the most informative feature (with the largest information gain). For GBDT, the information gain is usually measured by the variance after splitting. When O is training dataset on a fixed node of the decision tree. The variance gain of splitting feature j at point d for this node is define as

$$V_{j|O}(d) = \frac{1}{n_O}\left(\frac{(\sum_{\{x_i \in O: x_{ij} \le d\}} g_i)^2}{n^j_{l|O}(d)} + \frac{(\sum_{\{x_i \in O: x_{ij} > d\}} g_i)^2}{n^j_{r|O}(d)}\right)$$

For feature j, the decision tree algorithm selects $d_j^* = argmax_d V_j(d)$ and calculated the largest gain $V_j(d_j^*)$. (Guolin Ke, 2017)

Chapter 3. Research Method

3.1 Statistical Analysis

Data Extraction using R (4.1.2) and PostgreSQL (4.6.3). Data cleaning and analysis using python (version 3.9).

3.1.1 Database

The Medical Information Mart for Intensive Care (MIMIC-III) database was used as the data resource. (Johnson AEW, et al.) MIMIC-III is a single-center database covering 38,597 distinct adult patients admitted to the ICU in the Beth Israel Deaconess Medical Center in Boston from 2001 to 2012. MIMIC-IIII integrates comprehensive clinical data and makes them accessible to researchers

worldwide under data use agreement. I have obtained permission after application and completion of the course and test. I established and validated the prediction models using the retrospectively extracted data in MIMIC-III

3.1.2 Data Extraction and Exploratory Data Analysis

Age was calculated as the age at each admission. Height at the time of admission was used, and inches were converted to centimeters. COPD was marked as Diagnosed/Non-diagnosed according to the hospitalization number match. SAPS II and OASIS were extracted according to https://github.com/MIT-LCP/mimic-code and matched to hospitalization number. The average values of Vt, Ve, HR, MBP, RR, Pimax, and PaCO2 were calculated within 24 hours from the time of extubation. GCS, SpO2, and FiO2 were extracted as the latest values to collect data just before extubation.

Before all calculations, outliers were extracted by removing data from each variable. Female/male and COPD Diagnosed/Non-diagnosed were coded as one-hot-encoding. All rows with missing values were deleted to preserve the meaning of the original data. PaCO2 had the most missing values.

The P-value between the independent variable and the dependent

variable was calculated using statsmodel package of python.

3.1.3 Characteristic Analysis

Of the total 16 variables, 2 were categorical and 14 were continuous variables. Each variable was analyzed by dividing it into two groups: extubation failure and extubation non-failure. The total number of cases in the extubation failure group is 215, and the total number of cases in the extubation non-failure group is 532.

Categorical variables of characteristics were expressed as counts. Continuous variables of characteristics were expressed as mean (SD). The categorical variables are gender (female/male) and COPD (diagnosed/non-diagnosed).

In extubation failure, 84 cases were female, and 131 cases were male (female: male ratio was 0.64). In extubation non-failure, 121 cases were female, and 411 cases were male (female: male ratio was 0.29). In the extubation failure group, the female: male ratio was higher than that of extubation non-failure, and it was found that gender influenced extubation failure. In extubation failure, COPD was diagnosed in 27 cases and non-diagnosed in 188 cases (diagnosed: non-diagnosed ratio was 0.14). In extubation non-failure, COPD was diagnosed in 72 cases and non-diagnosed in 460 cases (diagnosed: non-diagnosed ratio was 0.15). There was no

significant difference between the extubation failure group and the extubation non-failure group.

The categorical variables are age, height, GCS (Glasgow coma scale), SAPS II (Simplified Acute Physiology Score II), OASIS(Outcome and Assessment Information Set), Vt(Tidal Volume), Ve(Minute Ventilation), HR(Heart Rate), MBP(Mean Blood Pressure), RR(Respiratory Rate), Pimax(Maximal Inspiratory Pressure), SpO2(Saturation of percutaneous oxygen), FiO2(Fraction of inspired oxygen), PaCO2(Carbon dioxide pressure in arterial blood).

The age average of the extubation failure group was 71.4 years, and the average of the extubation non-failure group was 64.8 years. Old age could affect extubation failure.

The height of the average of the extubation failure group was 169, and the average of the extubation non-failure group was 173. The average height of the non-failure group was greater.

GCS of the average of the extubation failure group was 2.8, and the average of the extubation non-failure group was 3.48. The average GCS of the failure group was smaller than non-failure group.

SAPS II of the average of the extubation failure group was 31.5, and the average of the extubation non-failure group was 36.2. OASIS of the average of the extubation failure group was 33.1, and the

average of the extubation non-failure group was 35.0. These two Severity Scores tended to be slightly higher in the extubation non-failure group.

In ventilation factors, which include Vt, Ve, HR, MBP, RR, Pimax, SpO2, FiO2 and PaCO2. There was no significant difference between the two groups and showed a similar trend.

In the p-value evaluation of all variables, according to Table 1, GCS is only variable has p-value less than 0.05.

Table 1 Characteristics between Extubation Failure and Extubation Non-failure

Variable	Extubation Failure n=215	Extubation Non-failure n=532	p-value n = 747
Demographic			
Age (years), mean(SD)	71.4 (47.3)	64.8 (30.8)	0.078
Gender (Females/Male)	84/131	121/411	0.125
Height (cm), mean(SD)	169(10.6)	173 (9.72)	0.084
GCS(Glasgow coma scale), mean(SD)	2.8 (2.29)	3.48 (2.35)	0.000
Relevant Diagnosis			
COPD (Diagnosed/Non-diagnosed)	27/188	72/460	0.157
Severity Score			
SAPS II, mean(SD)	31.5 (14.5)	36.2 (16.1)	0.265
OASIS, mean(SD)	33.1 (8.16)	35.0 (16.1)	0.637
Ventilation factors			
Vt (tidal volume), mean(SD)	501 (79.1)	512 (89.4)	0.705

Ve (minute ventilation), mean(SD)	9.10 (1.99)	10.0 (2.14)	0.213
HR (Heart Rate), mean(SD)	84.3 (12.5)	84.5 (15.2)	0.269
MBP (Mean Blood Pressure), mean(SD)	76.8 (8.67)	77.1 (9.68)	0.022
RR (Respiratory Rate), mean(SD)	18.2 (3.69)	19.2 (3.84)	0.110
Pimax (cmH2O), mean(SD)	18.7 (4.88)	18.6 (5.10)	0.353
SpO2, mean(SD)	97.8 (2.10)	97.7 (2.24)	0.082
FiO2, mean(SD)	47.6 (12.8)	47.0 (12.4)	0.646
PaCO2, mean(SD)	9.67 (1.92)	9.65 (1.57)	0.710

3.2 Model Design, fitting, hyperparameter

Models are developed using Python scikit-learn packages.

Model fitting is performed using training data which split 0.8 from cleansed dataset. Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost and Light GBM models were fitted to train data to obtain the best model through GridSearchCV using 5-fold validation. The best hyperparameter values of the model used for the model fitting process. The hyperparameter values of each best model are as follows. Logistic Regression is {'C': 0.1, 'class_weight': None, 'penalty': 'l2'}. KNN is {'n_neighbors': 5, 'weights': 'distance'}. SVM is { ‘kernel’ : ‘linear’ , ‘C’ =5}. Decision Tree is {'max_depth': 7, 'max_features': 0.8}. Random Forest is {'max_depth': 8, 'max_features': 0.8, 'n_estimators': 50}. XGBoost is {'learning_rate': 0.5, 'max_depth': 5, 'max_features': 0.7,

'n_estimators': 200, 'reg_lambda': 0.05}. Light GBM is {'learning rate': 0.0001, 'max_depth': 5, 'max_features': 0.8, 'n_estimators': 200, 'num_leaves': 15}.

Chapter 4. Results

4.1 Model Results

4.1.1 Model performance

The models were evaluated by sensitivity, specificity, accuracy, AUROC, and F1 score.

Precision is the number of correctly classified positive examples divided by the number of examples labeled by the system as positive. Recall(specificity) is the number of correctly classified positive examples divided by the number of positive examples in the data. Fscore is a combination of the above.

Table 2 Confusion matrix for binary classification

Data class	Classified as Positive	Classified as Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The correctness of a classification can be evaluated by computing

the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and examples that either were incorrectly assigned to the class (false positives) or that were not recognized as class examples (false negatives). These four counts constitute a confusion matrix for the case of the binary classification.

Table 3 Measure for binary classification using the notation of table 2

Measure	Formula	Evaluation Focus
Accuracy	$\frac{TP+TN}{TP+FN+FP+TN}$	Overall effectiveness of a classifier
Precision	$\frac{TP}{TP+FP}$	Class agreement of the data labels with the positive labels given by the classifier
Recall (Sensitivity)	$\frac{TP}{TP+FN}$	Effectiveness of a classifier to identify positive labels
Fscore	$\frac{(\beta^2+1)TP}{(\beta^2+1)TP+\beta^2FN+FP}$	Relations between data's positive labels and those given by a classifier
Specificity	$\frac{TN}{FP+TN}$	How effectively a classifier identifies negative labels
AUC	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$	Classifier's ability to avoid false classification

table 3 presents the most often used measures for binary classification based on the values of the confusion matrix. AUC (Area Under the Curve), 3 captures a single point on the Reception Operating Characteristic curve. However, we present Fscore's

properties because of its extensive use in text classification. (Marina Sokolova, 2009)

A binary classification problem is common in the medical field, often using sensitivity, specificity, accuracy, negative and positive predictive values as measures of performance of the binary predictor. In computer science, a classifier is usually evaluated with precision (positive predictive value) and recall(sensitivity). As a single summary measure of a classifier’ s performance, F1 score, defined as the harmonic mean of precision and recall, is widely used in the context of information retrieval and information extraction evaluation since it possesses favorable characteristics, especially when the prevalence is low. (Takahashi, 2021)

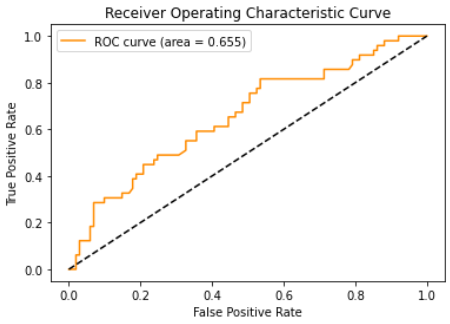
The best models were trained through hyperparameters of each method were evaluated as test data. The accuracy, sensitivity, specificity, F1 score, and AUROC of each model were shown in Table 4.

Table 4 Accuracy, Sensitivity, specificity, F1 score and AUROC for predict extubation failure

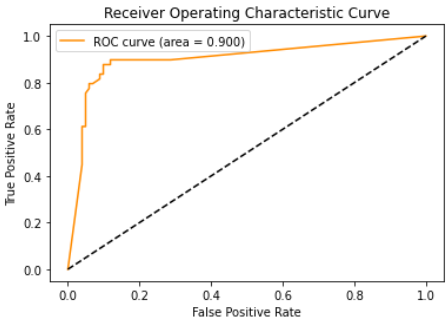
	Logistic Regression	KNN	SVM	Decision Tree	Random Forest	XGBoost	Light GBM
Accuracy	0.687	**0.893**	0.733	0.873	**0.893**	**0.900**	**0.907**
Sensitivity	0.184	0.796	0.347	0.694	0.755	**0.837**	**0.816**
Specificity	**0.931**	**0.941**	**0.921**	**0.960**	**0.960**	**0.931**	**0.950**

F1 Score	0.277	0.830	0.459	0.782	0.822	**0.845**	**0.851**
AUROC	0.655	**0.900**	0.620	0.850	**0.965**	**0.966**	**0.966**

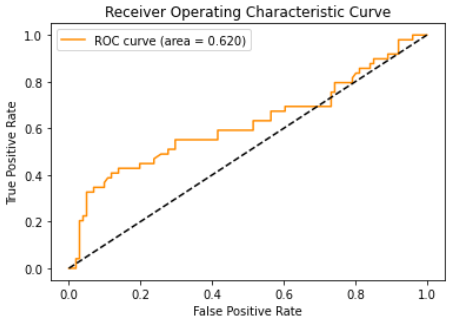
The accuracy of each model evaluated using the test set was as follows. Logistic Regression was 0.687, KNN was 0.893, SVM was 0.733, Decision Tree was 0.873, Random Forest is 0.893, XGBoost is 0.900 and Light GBM was 0.907. XGBoost and Light GBM show outstanding accuracy which exceeded 0.9. random forest and KNN also show high accuracy comparable to that. The sensitivity of XGBoost was 0.837 and Light GBM was 0.816. Specificity of All models were over 0.9. The specificity of XGBoost was 0.931 and Light GBM was 0.950. F1 Score of XGBoost is 0.845 and Light GBM is 0.851. XGBoost and Light GBM showed similar values where it was difficult to distinguish the difference between the two models. Figure 5 is a visual representation of the ROC curve. AUROC of KNN was 0.9, Random Forest was 0.965, Light GBM was 0.966 and XGBoost was 0.966.



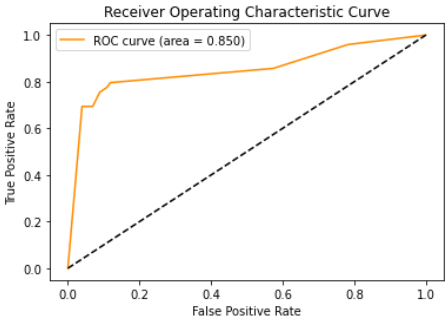
Logistic Regression



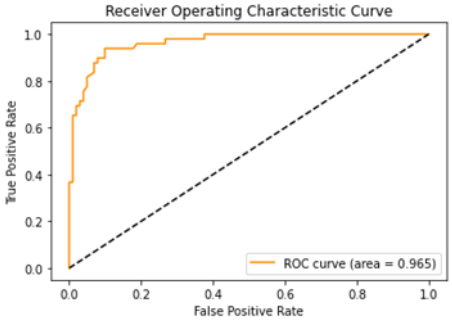
KNN



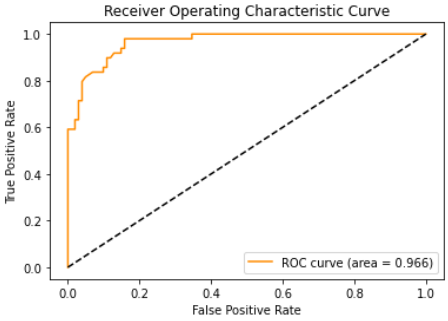
SVM



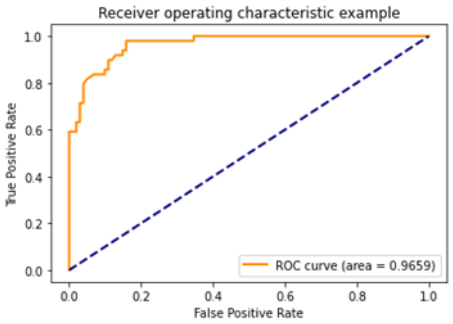
Decision Tree



Random Forest



XGBoost

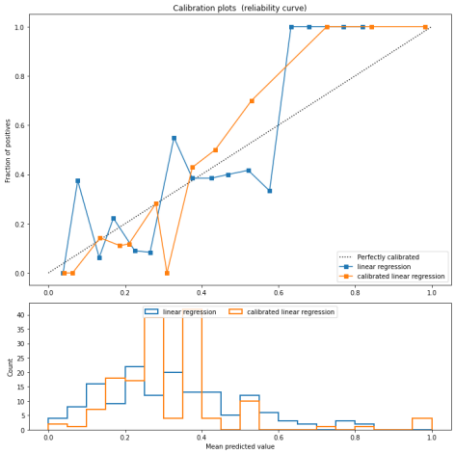


Light GBM

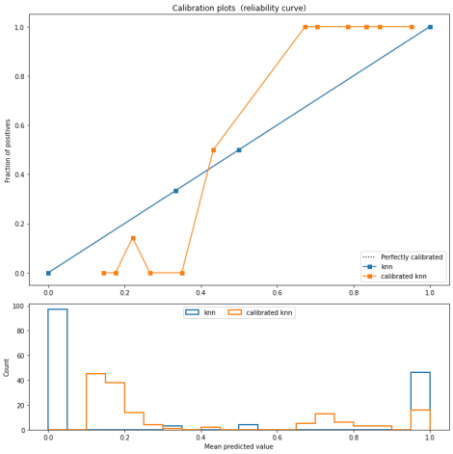
Figure 5 Receiver operating characteristic (ROC) curves of the seven models

4.1.2 Calibration plot

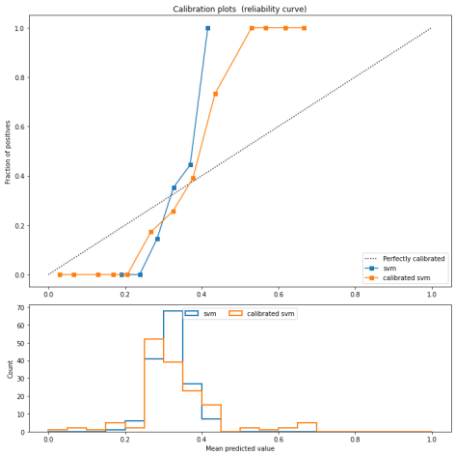
The best possible method of measuring the performance of a classifier's probability prediction on a dataset is using the calibration curve which is also referred to as a standardized curve. The calibration curve is created as Figure 6. (B. C. Wallace and I. J. Dahabreh, 2012).



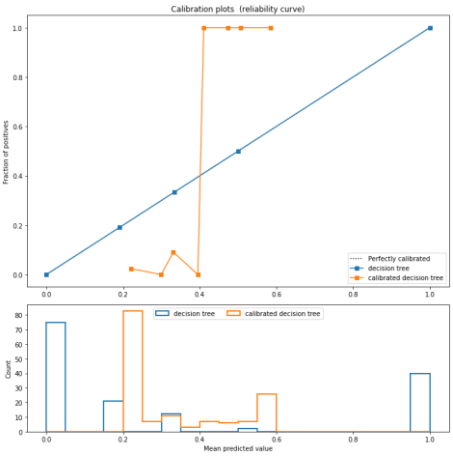
Logistic Regression



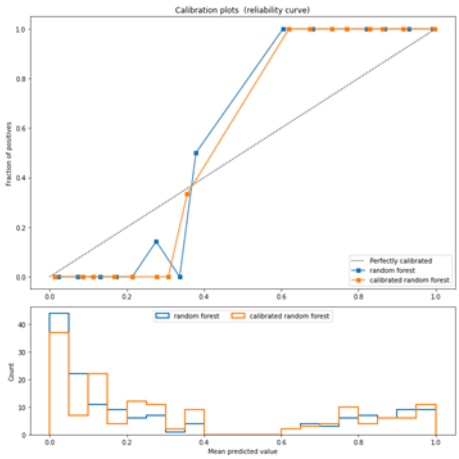
KNN



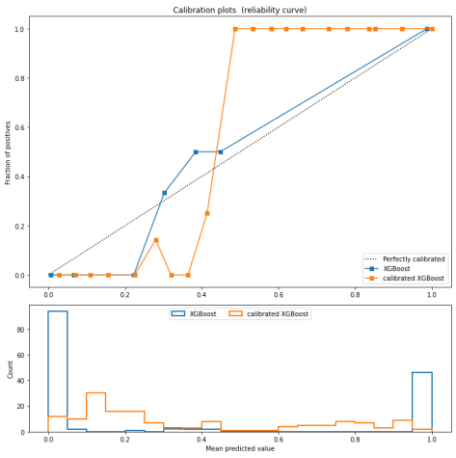
SVM



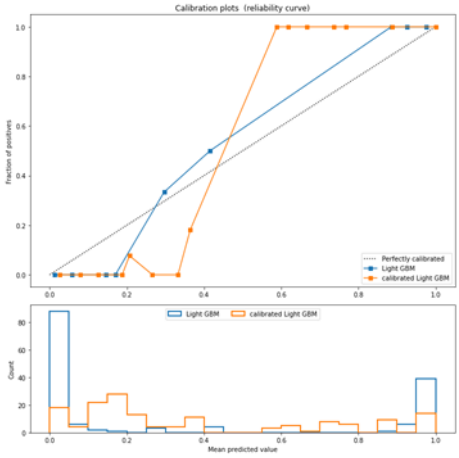
Decision Tree



Random Forest



XGBoost



Light GBM

Figure 6 Calibration Plots of the seven models

The dotted line (standardized curve) shows the perfect calibrated curve. This research tested the various classifier models such as Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost and Light GBM.

The graph illustrates the deviation of seven different classifier models from the non-calibrated curve and calibrated curve. KNN, Decision tree, XGBoost, Light GBM showed better non-calibrated curve than the calibrated curve. Logistic regression, SVM was slightly improved in the calibrated model, but this model had low predict-probability. Random forest was almost similar between the non-calibrated curve and calibrated curve. XGBoost and Light GBM is good to predict probability in the non-calibrated curve.

4.1.3 Best Model Selection

XGBoost and Light GBM models were showed outstanding performance and better predict probability than other models. but the purpose of the classification model is to distinguish between true positive rate (Sensitivity) and true negative rate (Specificity). Also in the medical field, it can be important to ensure a high degree of specificity in diagnosis. Considering the purpose of a test in certain populations requires careful consideration of both sensitivity and specificity. This helps both healthcare providers and patients to make the best decisions about testing and treatment. The higher model's specificity, the less often it will incorrectly find a result it is not supposed to.

Among models with similar accuracy and AUROC, the model that

classifies sensitivity and specificity in a balanced way could be better than the highest sensitivity model. Therefore, in this research, XGBoost can be interpreted as the best classification model.

4.2 Feature Importance

In this research, data included in previous studies and variables traditionally considered to be related to extubation failure were included. And, in order to be easily applied in external validation or other institutions in the future, I tried to exclude as much as possible the variables that can be Institution-specific, and I tried to develop a model that can achieve the best performance with simple features. The models showing the best performance were XGBoost and Light GBM. It was difficult to distinguish between superiority and inferiority, and the variables that affected each model were ranked by the model using SHAP, and the learning variables were compared and analyzed. And selected the top 5 variables.

Comparing the importance of variables used in model training can distinguish the influence of variables that are strongly influenced and those that do not, among variables expected to be actually related. Therefore, through this analysis, we can confirm how much

the variables that were thought to be statistically or related in past studies helped predictive learning.

XGBoost was selected as the best model in this research. According to Figure7, Ve, GCS, Height, Vt, OASIS, and Age were the top 5 variables in the XGBoost model. The most impact variable was Ve. In XGBoost, the variables considered significant in previous research such as Pimax, COPD, and SpO2, had little effect on model output.

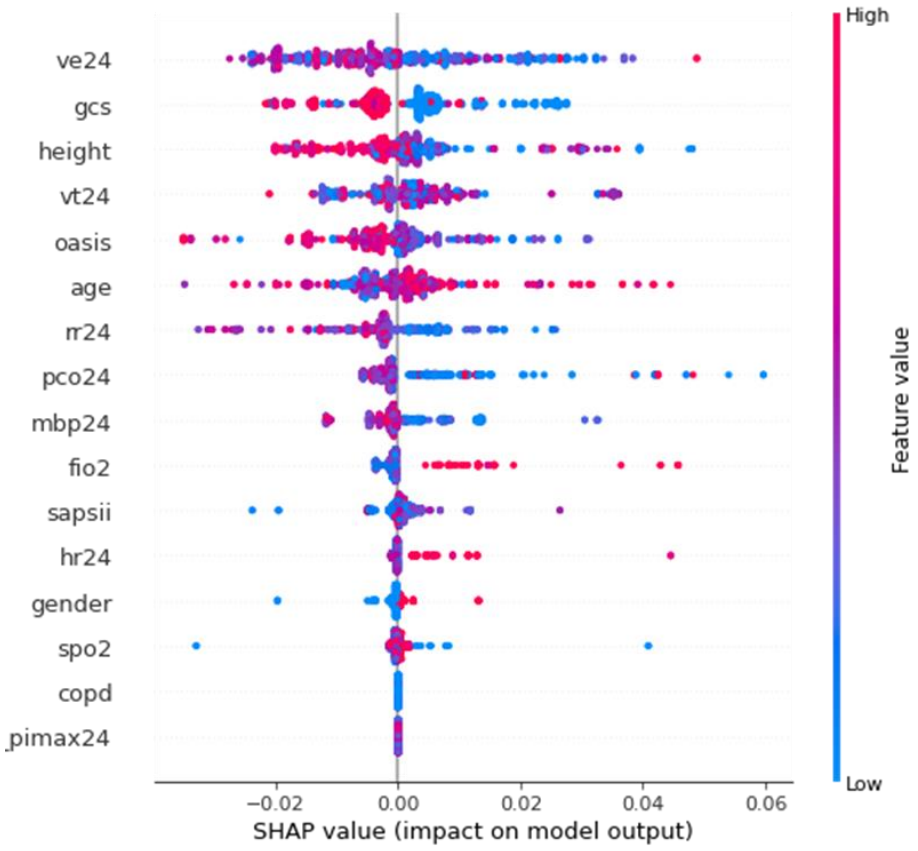


Figure 7 SHAP value of XGBoost

In SHAP of Light GBM(Figure 8), The top 5 variables were Ve, OASIS, HR, SpO2, and MBP. Ve was the most impact variable in Light GBM, like XGBoost. OASIS was considered an important variable in these two models. SpO2 was the low importance in XGBoost model, but in Light GBM, it was an important variable. COPD was the lowest level important variable on both models.

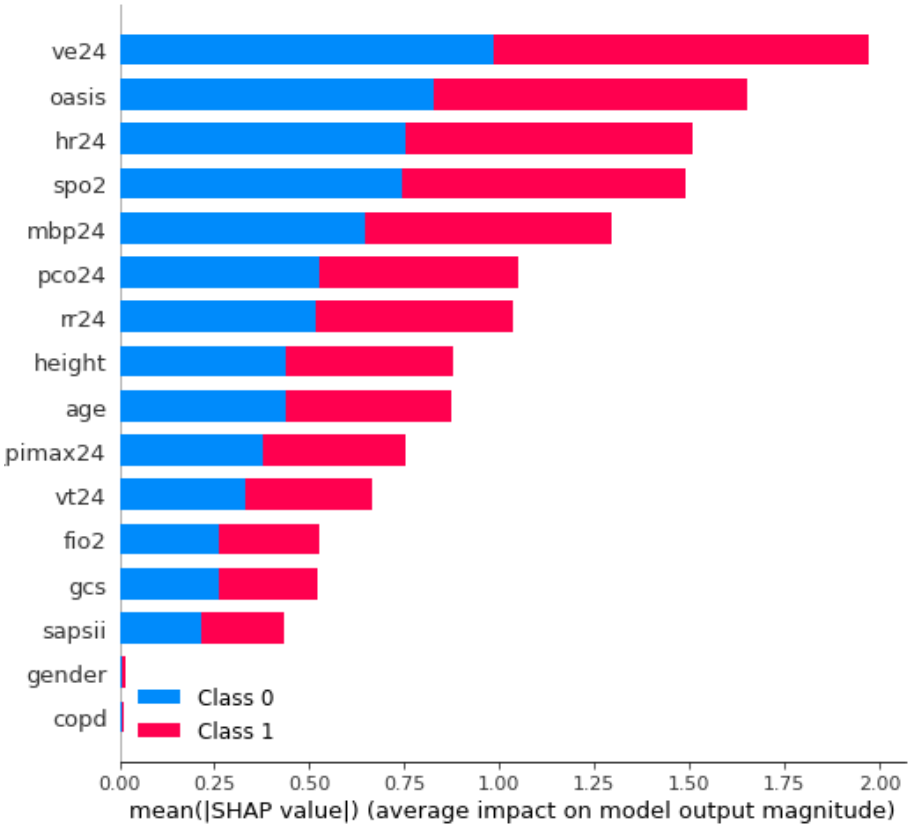


Figure 8 SHAP value of Light GBM

Chapter 5. Conclusion

5.1 Summary and Implications

Weaning failure is an important issue in mechanical ventilation. To assess the optimal timing of weaning and extubation, various weaning indices were used such as the rapid shallow breathing index (RSBI), maximal inspiratory pressure (Pimax), airway occlusion pressure in the first 100 ms (P0.1), and P0.1/Pimax. However, the reintubation rate after extubation failure is 6% to 47%. Therefore, when deciding on extubation, it is necessary to consider various factors along with the process of reducing dependence on mechanical ventilation, and more accurate indicators to improve the current failure rate should be developed.

In this research, extubation cases of adult patients who performed mechanical ventilation in the MIMIC-III database were classified into extubation failure and extubation non-failure groups, and Logistic regression, KNN, SVM, Decision Tree, Random Forest, XGBoost and Light GBM were trained and tested. Then, select the best model that more accurately predicts extubation failure by comparing and analyzing various classification models.

As a result of this research, the accuracy of KNN, Decision Tree,

XGBoost, and Light GBM were 0.893, 0.893, 0.900, and 0.907. Sensitivity, Specificity, F1 Score, AUROC, and Calibration Plot were compared to select the Best Model among XGBoost and Light GBM. XGBoost was selected as the best model because Sensitivity and Specificity were balanced at 0.837 and 0.931.

In addition, the feature importance of XGBoost and Light GBM and Top 5 variables were compared. Ve and Oasis were observed in both models, and the top 5 variables of XGBoost were Ve, GCS, Height, Vt, and OASIS. COPD onset and gender, which were traditionally considered significant, did not have a significant effect on the learning of the model.

This research was the only paper that compares and analyzes extubation failure with various machine learning classifiers. This research designed more subjects were recruited to increase power, selected minimized variables to expand applicability. Subject's indications which used for model training were not disease-specific, and to conduct comparative analysis of various machine learning methods. Logistic Regression, KNN, SVM, Decision Tree, Random Forest, XGBoost and Light GBM selected to find best classification model.

This research showed high model performance, and the Accuracy, Sensitivity, Specificity, F1 score and AUROC of the best model

which trained XGBoost algorithm selected in this research were 0.900, 0.837, 0.931, 0.845, 0.966. It had shown outstanding results in medical classification models. This model could be used widely for improving extubation failure.

Model using the time series data might be difficult to apply to the real world depending on institutions. In addition, too many variables could increase the complexity of the model and increase the amount of computation cost which could become a hurdle that is difficult to use in the real world. In this research, to overcome such limitations, this research was trained and evaluated using only 13 variables and data within 24 hours to determine the patient's extubation. Nevertheless, the result of AUC was higher than that of Chung's research using time series data as neural network.

Therefore, this research has good scalability and can be used in practice to relatively accurately classify whether or not extubation fails or not, using patient data within 24 hours before the medical staff decides on extubation of the patient. This model allows the medical staff to determine the timing of extubation and it could help medical decisions.

5.2 Research Limitation and Future Plans

There had limitations in this research. First, models were retrospectively studied based on a single-center database. Thus, further prospective or multicenter research is needed to evaluate the generalization of models and predictors. Second, there were missing data in this research. Due to missing data, it was difficult to predict the exact subject's condition at a specific time point, and variables with many missing data among potential variables could not be included in the model design. In this research, to preserve the meaning of variables, missing data were eliminated without imputation or scaling. However, if using imputation of missing values in future research, more data could be utilized, and different insights from this research could be obtained. Third, external validation has not been employed in this research. In future research, If model shows good performance in external validation, the variables and models used in this research could get scalability. Forth, Despite the high model performance, uniform results were not obtained in the calibration plot. Rather, most of the models showed more stable predict probability in the non-calibrated plot. There is a need to build datasets from various perspectives and test them in future studies.

Reference

B. C. Wallace and I. J. Dahabreh. (2012). Class Probability Estimates are Unreliable for Imbalanced Data (and How to Fix Them). 2012 IEEE 12th International Conference on Data Mining pp. 695-704.

Brown CV, Daigle JB, Foulkrod KH, Brouillette B, Clark A, & Czysz C. (2011). Risk factors associated with early reintubation in trauma patients: a prospective observational study. J Trauma 2011. Jul;71(1):37-41, discussion 41-42 10.1097.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Association for Computing Machinery, New York, NY, USA, 785–794.

Chung WC, Sheu CC, Hung JY, Hsu TJ, Yang SH, & JR., T. (n.d.). Novel mechanical ventilator weaning predictive model. Kaohsiung J Med Sci. 2020 Nov;36(11).

Epstein SK, & RL., C. (1998). Independent effects of etiology of failure and time to reintubation on outcome for patients failing extubation. Am J Respir Crit Care Med 1998;158:489-93.

Epstein, S., Ciubotaru, R., & Wong, J. (1997). Effect of failed extubation on the outcome of mechanical ventilation. Chest 1997. Jul;112(1):186-192 10.1378/chest.112.1.186.

Esteban, A., Frutos, F., Tobin, M., Alía, I., Solsona, J., & Valverdú, I. (1995). Spanish Lung Failure Collaborative Group A comparison of four methods of weaning patients from mechanical ventilation. N Engl J Med 1995. Feb;332(6):345-350.

Guolin Ke, Q. M.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

J-M. Boles, J. B.-B. (2007). Weaning from mechanical ventilation. European Respiratory Journal May 2007, 29(5) 1033-1056.

Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, & Ghassemi M. (n.d.). MIMIC-III, a freely accessible critical care database. *Sci Data. (2016) 3:160035. doi: 10.1038/sdata.2016.35.*

Kuilkarni AP, & Agarwal., V. (2008). Extubation failure in intensive care unit. Predictors and management. Indian J Crit Care Med. 2008;12(1), 4103/2-5229.

Kyu-Hyouck, K. (2012). Weaning from Mechanical Ventilation. Journal of Acute Care Surgery 2012 2:2 p44-48.

M.Tobin., & C.Manthous. (2017). Mechanical ventilation. Am J Respir Crit Care Med. 196(2):P3-P4.

Mahmood S, Alani M, Al-Thani H, Mahmood I, El-Menyar A, & R., L. (2014). Predictors of reintubation in trauma intensive care unit: qatar experience. Oman Med J. 2014;29(4):289-293. doi:10.5001/omj.2014.75.

Marina Sokolova, G. L. (2009). A systematic analysis of performance measures for classification tasks. (pp. 427-437). Information Processing & Management,Volume 45, Issue 4,.

Metnitz, P., Metnitz, B., Moreno, R., & al., e. (2009). Epidemiology of Mechanical Ventilation: Analysis of the SAPS 3 Database. Intensive Care Med 35, 816-825.

MJ., T. (2001). Advances in mechanical ventilation. N Engl J Med:344(26):1986-1996.

RL., M., & RP., C. (1996). Association between reduced cuff leak volume and postextubation stridor. Chest 1996. Oct;110(4):1035-1040 10.1378/chest.110.4.1035.

S., B. U., Souza. GF., Campos. ES., Farah. De. Carvalho. E., Fernandes. MG., & I., S. (2015). Maximum inspiratoryry pressure and rapid shallow breathing index as predictors of successful ventilator weaning. J Phys Ther SCi.2015;27(12):3723-3727.

Seymour CW, Martinez A, Christie JD, & BD., F. (2004). The outcome of extubation failure in a community hospital intensive care unit: a cohort study. Crit Care 2004. Oct;8(5):R322-R327 10.1186/cc2913.

Takahashi, K. Y. (2021). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. . Appl Intell.

Tan, P.-n. (2019). *Introduction to Data Mining 2nd edition.*

Wonsch.H. (2013). ICU occupancy and mechanical ventilator use in the

United States. Crit Care Med, 41:2712-9.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., . . . Zhou, Z.-H. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems. 14 (1): 1-37.

Zhu Yibing, Zhang Jin, Wang Guowei, Yao Renqi, Ren Chao, Chen Ge, . . . Qian, Y. (2021). Machine Learning Prediction Models for Mechanically Ventilated Patients: Analyses of the MIMIC-III Database. (p. 955). Frontiers in Medicine.