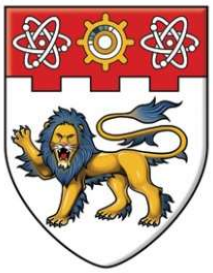

NANYANG TECHNOLOGICAL UNIVERSITY

The risk factor identification of in-hospital mortality based on heart failure patients in MIMIC II database

(BS6200: ESSENTIAL MACHINE LEARNING FOR BIOMEDICAL SCIENCE)

G2201034B

ZHANG ZHU

Contents

1. Introduction

Heart failure (HF), a syndrome due to a heart function disorder, is the terminal phase of all heart diseases. As a major cause of cardiovascular morbidity and mortality, HF has become an overwhelming threat to human health and social development. Despite the recent progress in diagnosis and evidence-based management, the outcomes concerning HF remain unsatisfactory. The main purpose of the project is to recognize the risk factors of in-hospital mortality among heart failure patients.

2. Methods

2.1. Dataset

Multiparameter Intelligent Monitoring in Intensive Care[1] (MIMIC II for short) database was collected aiming to develop and evaluate advanced Intensive Care Unit (ICU) patient monitoring systems that will substantially improve the efficiency, accuracy and timeliness of clinical decision making in intensive care. It contains detailed clinical information for each patient in their ICU days, including demographical characteristics, diagnosis, procedures, lab measurements, microbiology test, chart notes, prescriptions, comorbidity and so on.

2.2. Patient and feature selection

The patients diagnosed with heart failure and ≥15 years old at the time of ICU admission were included in the study. All sort of heart failure was collected for downstream analysis, encompassing congestive heart failure, systolic heart failure, diastolic heart failure, acute and chronic heart failure.

The candidate variables were decided based on relative research[2-4].

Using Python, demographic characteristics, vital signs and laboratory values data were extracted from the following tables in the MIMIC-II dataset: d_patients, ICD_9, chartevent, comorbidity_scores and labevent.

The subject IDs were used to identify distinct adult patients. The predictors included:
(a) Demographic information: age and gender;

(b) Comorbidities: cardiac arrhythmias, valvular disease, pulmonary circulation, peripheral vascular, hypertension, chronic pulmonary, diabetes uncomplicated and diabetes complicated;

(c) Vital signs: heart rate, mean blood pressure, respiratory rate, body temperature, saturation pulse oxygen (SPO2);

(d) Laboratory measurements: Creatinine, Urea nitrogen, Sodium, Chloride, Bicarbonate, Glucose, Anion gap, Platelets, Leukocytes, Erythrocyte mean corpuscular

hemoglobin concentration, Erythrocytes, Erythrocyte mean corpuscular volume, Erythrocyte distribution width, Magnesium, pH of Blood, Calcium, Carbon dioxide in Blood, Oxygen in Blood, Prothrombin time in Blood, Creatine kinase in Serum or Plasma, Lactate, Basophils/100 leukocytes in Blood, Neutrophils segmented/100 leukocytes in Blood, Lymphocytes/100 leukocytes in Blood, Monocytes/100 leukocytes in Blood, Eosinophils/100 leukocytes in Blood, Neutrophils band form/100 leukocytes in Blood.

2.3. Data processing

The information in labevent and chartevent was endoded by retieving item id in two dictionary files: d_chartitems and d_labitems.

Vital signs and laboratory variables were measured during the entire ICU stay.

For variable data with multiple measurements, the calculated mean value was included for analysis. The primary outcome of the study was in-hospital mortality,

defined as the vital status at the time of hospital discharge in survivors and non-survivors.

2.4. Outlier and missing value

Outlier was detected by boxplot and removed manually, followed by filling, which depend on distribution of variables.

Variables with missing data are commonplace in the MIMIC-II, however, eliminating patients with incomplete data can bias the study. Therefore, imputation is an important step in data preprocessing. All screening variables contained >25% missing values were removed. For normally distributed continuous variables, the missing values were replaced with the mean for the patient group. For skewed distributions, missing values were replaced with their median. There were no missing dichotomous variables in this study.

2.5. Statistical analysis

The Shapiro Wilk test was implemented for normalization test.

The normal distributed variables are described as mean $\pm$ SD, and non-normal distributed variables are described as median (Interquartile Range). For continuous variables, Student's t test or Wilcoxon rank sum test was applied to compare the significant difference between two groups, depending on their distributions. For categorical variables, Pearson Chi-Square test was applied to assess differences in proportions between the two groups.

2.6. Dimension reduction

Principal component analysis and t-SNE were used to reduce the dimension of data followed with visualization.

2.7. Model

Both classical and novel algorithms participated in this study. They can be generally divided into 3 groups:

(1) Classical or basic algorithms

K-nearest neighbor, Decision tree, Support Vector Machine and Multilayer perceptron.

(2) Ensemble algorithms (tree-based)

Random Forest takes the majority vote of decision trees.

AdaBoost: a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

Gradient boosting: as its name implies, it's the combination of gradient and boosting. It uses gradient descent to minimize the loss function, whereby the best hyperparameter of tree can be searched.

Extreme Gradient Boosting: a specific implementation of the Gradient Boosting method which uses more accurate approximations to find the best tree model.

(3) Algorithms combined with feature selection

To further investigate the impact of feature selection, 2 combinations of Recursive Feature Elimination and Gradient Boosting or Extreme Gradient Boosting were incorporated in analysis.

The intent of this design is to compare the performance on different angle of view.

2.8. Training and testing

The entire data set was split into training set (75%, n = 1194) and test set (25%, n = 399). 10 classifiers were trained on training set and evaluate on both of training set and test set respectively.

To equip each model with optimal hyperparameters, RandomizedSearchCV was applied, which implements a randomized search over parameters, where each setting is sampled from a distribution over possible parameter value. The searching outcomes will be applied for each model.

2.9. Model evaluation

The scrupulous assessments were deployed for each classifier. Receptor Operation

Characteristic Curve, Calibration curve, f1-score and cross validation score. The numeric measurements including Area Under Curve, weighted f1-score and mean of cross validation were calculated for model selection.

2.10.Feature selection

We can obtain feature importance information directly from model's attributes. The top 10 features will take part in downstream analysis. However, the number of selected features of RFECV algorithms is uncertain. To make sure each model contributes to the final result equally, the permutation feature importance was introduced.

Apart of feature importance, permutation importance is another measurement to evaluate the feature's impact on model. The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled[5]. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature[6].

2.11.Shap value and feature explanation

SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model, which assigns each feature an importance value for a particular prediction[7]. Originating in game theory, SHAP values partition the prediction result of every sample into the contribution of each constituent feature value. This is done by estimating differences between models with subsets of the feature space. By averaging across samples, SHAP values estimate the contribution of each feature to overall model predictions.[8]

Some studies in biomedical domain have already introduced shap value to interpret the impact of features, including clinical data or electronic health records (EHC) investigation[8-10].

In this study, shap value and the function derived from it are used for explain how the features impact the result. To see how these indicators influence the in-hospital mortality, Partial dependence plots (PDP), which is based on shap value were created for each selected variable.

3. Results

3.1. Data summary

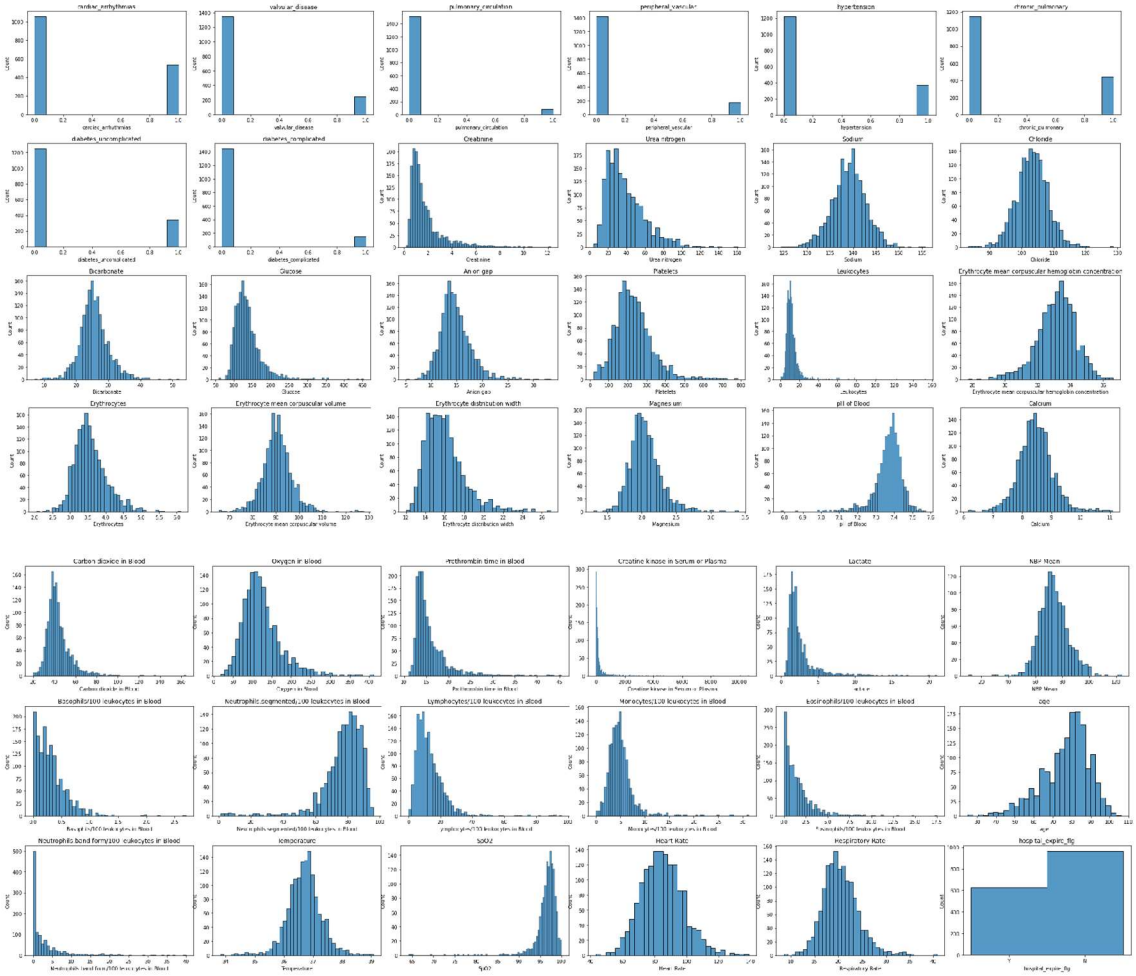

Figure 1 . Histogram for each variable

3.2. Statistical analysis

Table 1. Statistical analysis for continuous variable

	Total (n = 1593)	**Survive (n = 967)**	**Expire (n = 626)**	**P Value**
Creatinine	1.32 (0.91-2.05)	1.22 (0.86-1.89)	1.49 (1.01-2.29)	< 0.001
Urea nitrogen	33.5 (23.25-49.67)	30.03 (21.41-43.86)	40.08 (28.02-57.85)	< 0.001
Sodium	139.2 (136.89-141.34)	139.1 (137.0-141.0)	139.42 (136.58-142.0)	0.086
Chloride	103.2 (100.08-106.29)	102.83 (99.99-105.76)	104.0 (100.25-107.48)	< 0.001
Bicarbonate	25.25 (22.95-27.81)	25.8 (23.79-28.35)	24.08 (21.53-26.73)	< 0.001
Glucose	129.58 (113.25-150.0)	125.91 (110.82-146.65)	134.03 (118.81-154.15)	< 0.001
Anion gap	14.26 (12.78-16.19)	13.97 (12.57-15.53)	14.87 (13.19-17.27)	< 0.001
Platelets	217.13 (165.18-278.32)	225.63 (171.94-286.48)	203.68 (151.72-262.36)	< 0.001
Leukocytes	10.76 (8.28-13.52)	10.39 (8.03-12.58)	11.74 (8.83-15.33)	< 0.001

Erythrocyte mean corpuscular hemoglobin concentration	33.25 (32.47-33.93)	33.29 (32.5-33.97)	33.17 (32.35-33.85)	0.065
Erythrocytes	3.5 (3.26-3.79)	3.52 (3.3-3.79)	3.45 (3.21-3.81)	0.002
Erythrocyte mean corpuscular volume	90.4 (86.88-93.88)	89.9 (86.87-93.28)	90.89 (86.96-94.56)	0.001
Erythrocyte distribution width	15.65 (14.52-16.91)	15.53 (14.45-16.67)	15.8 (14.69-17.41)	< 0.001
Magnesium	2.02 (1.9-2.16)	2.0 (1.89-2.13)	2.05 (1.94-2.23)	< 0.001
pH of Blood	7.37 (7.34-7.41)	7.38 (7.36-7.41)	7.37 (7.31-7.4)	< 0.001
Calcium	8.47 (8.12-8.8)	8.5 (8.18-8.8)	8.37 (7.99-8.77)	< 0.001
Carbon dioxide in Blood	42.33 (37.5-46.88)	42.75 (38.12-46.13)	41.42 (36.41-47.98)	0.125
Oxygen in Blood	115.13 (94.25-140.75)	115.13 (96.0-143.41)	115.13 (91.36-137.26)	0.064
Prothrombin time in Blood	14.57 (13.53-16.75)	14.4 (13.38-16.12)	14.93 (13.7-17.58)	< 0.001
Creatine kinase in Serum or Plasma	98.91 (54.67-197.5)	98.91 (53.6-180.25)	98.91 (59.12-228.91)	0.022
Lactate	1.94 (1.58-2.53)	1.94 (1.52-2.13)	2.07 (1.7-3.3)	< 0.001
Basophils/100 leukocytes in Blood	0.25 (0.13-0.4)	0.25 (0.17-0.43)	0.22 (0.09-0.36)	< 0.001
Neutrophils.segmented/100 leukocytes in Blood	79.65 (73.4-85.0)	79.65 (73.07-84.04)	80.35 (74.12-86.2)	0.001
Lymphocytes/100 leukocytes in Blood	10.92 (7.46-15.88)	10.92 (8.26-16.65)	9.95 (6.28-14.33)	< 0.001
Monocytes/100 leukocytes in Blood	4.66 (3.58-5.73)	4.66 (3.71-5.9)	4.64 (3.29-5.48)	< 0.001
Eosinophils/100 leukocytes in Blood	1.02 (0.45-1.95)	1.1 (0.6-2.2)	0.9 (0.27-1.52)	< 0.001
Neutrophils.band form/100 leukocytes in Blood	1.0 (0.11-2.57)	1.0 (0.0-1.79)	1.0 (0.5-3.67)	< 0.001
Temperature	36.69 (36.4-36.98)	36.69 (36.43-36.95)	36.69 (36.33-37.03)	0.578
SpO2	97.0 (96.1-97.75)	97.0 (96.35-97.94)	97.0 (95.66-97.31)	< 0.001
Heart Rate	84.76 (76.1-92.58)	83.68 (75.12-91.05)	85.47 (78.2-95.86)	< 0.001
Respiratory Rate	20.4 (18.08-22.44)	20.15 (18.0-22.02)	20.45 (18.26-22.91)	0.001
NBP Mean	73.67 (69.48-76.92)	73.67 (70.71-79.13)	73.67 (66.73-73.67)	< 0.001
age	79.0 (70.0-86.0)	79.0 (70.0-86.0)	79.0 (69.0-85.0)	0.548

Most of continuous variables have p value less than 0.05, which means for these variables, there are significant difference between 2 groups.

Table 2. Statistical analysis for categorical variable

	Total (n=1593)	Survive (n=967)	Expire (n=626)	P Value
sex				0.967
Male	824 (51.73%)	508 (52.53%)	316 (50.48%)	
Female	769 (48.27%)	459 (47.47%)	310 (49.52%)	
cardiac_arrhythmias				0.941
Yes	536 (33.65%)	312 (32.26%)	224 (35.78%)	
No	1057 (66.35%)	655 (67.74%)	402 (64.22%)	
valvular_disease				0.94
Yes	244 (15.32%)	158 (16.34%)	86 (13.74%)	
No	1349 (84.68%)	809 (83.66%)	540 (86.26%)	
pulmonary_circulation				0.99
Yes	79 (4.96%)	49 (5.07%)	30 (4.79%)	
No	1514 (95.04%)	918 (94.93%)	596 (95.21%)	
peripheral_vascular				0.959
Yes	170 (10.67%)	109 (11.27%)	61 (9.74%)	
No	1423 (89.33%)	858 (88.73%)	565 (90.26%)	
hypertension				0.92
Yes	370 (23.23%)	240 (24.82%)	130 (20.77%)	
No	1223 (76.77%)	727 (75.18%)	496 (79.23%)	
chronic_pulmonary				0.92
Yes	444 (27.87%)	286 (29.58%)	158 (25.24%)	
No	1149 (72.13%)	681 (70.42%)	468 (74.76%)	
diabetes_uncomplicated				0.945
Yes	345 (21.66%)	220 (22.75%)	125 (19.97%)	
No	1248 (78.34%)	747 (77.25%)	501 (80.03%)	
diabetes_complicated				0.881
Yes	146 (9.17%)	103 (10.65%)	43 (6.87%)	
No	1447 (90.83%)	864 (89.35%)	583 (93.13%)	

3.3. Dimension reduction

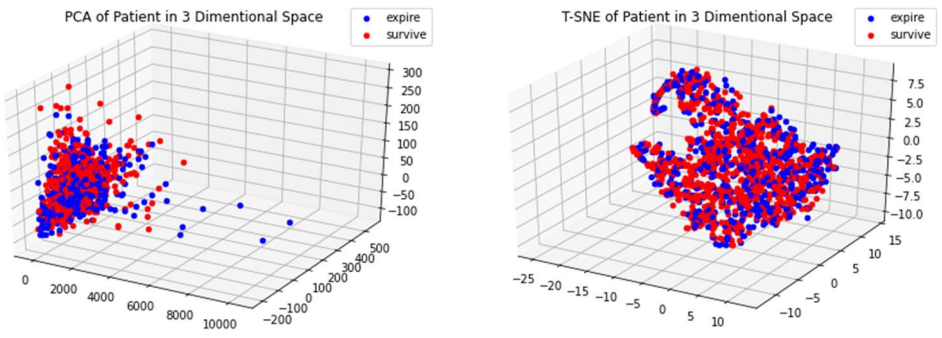

Figure 2 . Dimension reduction visualization derived from PCA and t-SNE

There is no perspicuous clustering tendency within each group.

3.4. Tree visualization

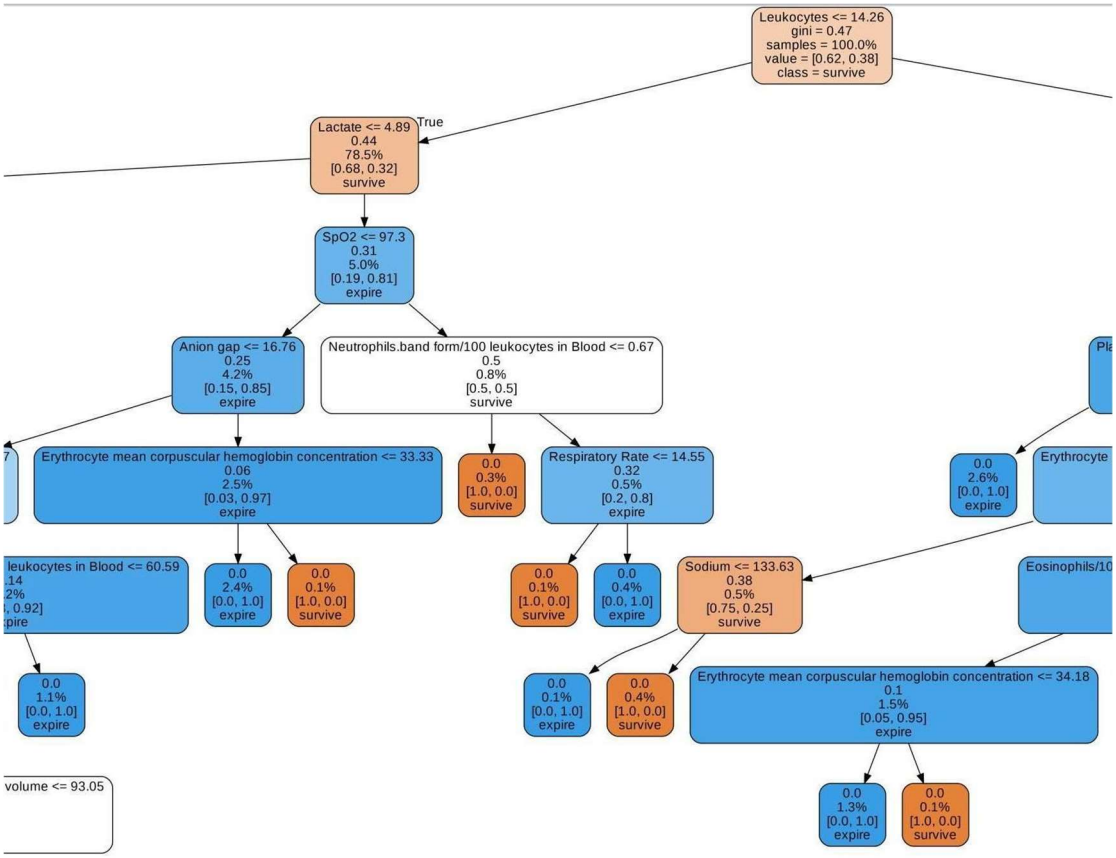

Figure 3. A part of single tree in random forest

There are mainly 2 hues of colors, orange and blue, which represent survive and expire respectively. The higher the saturation, the higher the proportion of corresponding patients. The first split is based on the level of leukocytes, and its Gini index is equal to 0.47. This node contains 62% of survive patients, so it is colored in pale orange.

Along with the splitting, Gini index decrease gradually and the saturation of nodes increase, which means the purity of nodes is raising. Finally, the tree stops splitting when the node is absolutely pure. This graphical visualization provides a perceptual intuition of tree-splitting algorithms.

3.5. Learning curve

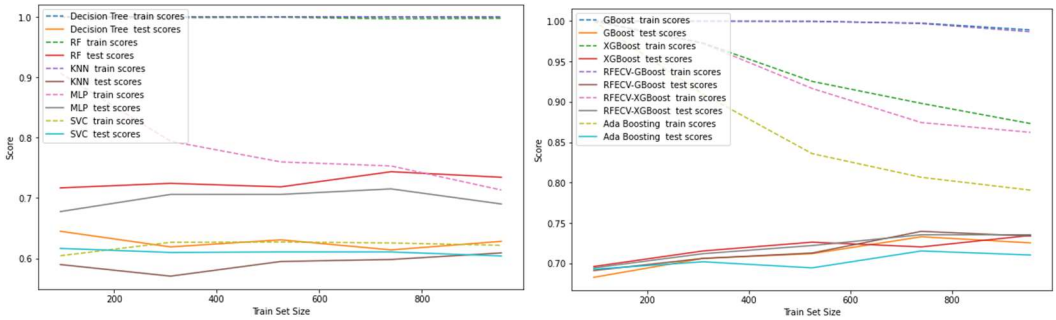

Figure 4. Learning curve for each model

3.6. ROC

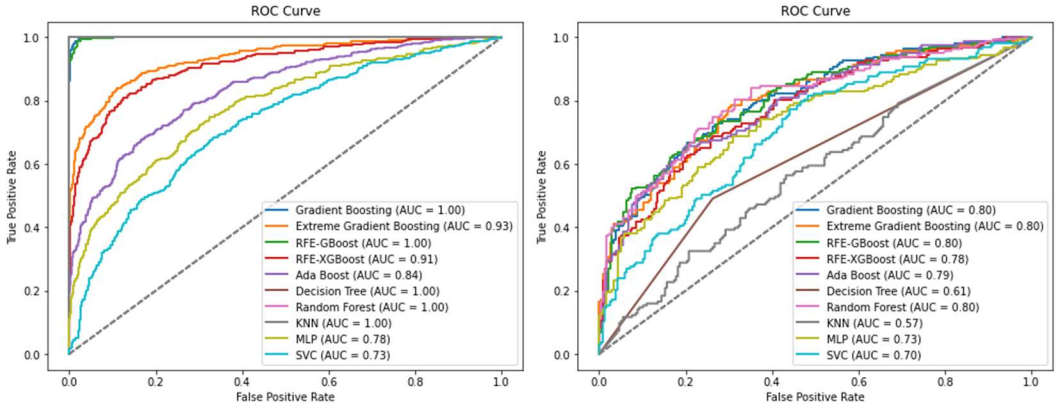

Figure 5. ROC curve for each model

3.7. Calibration curve

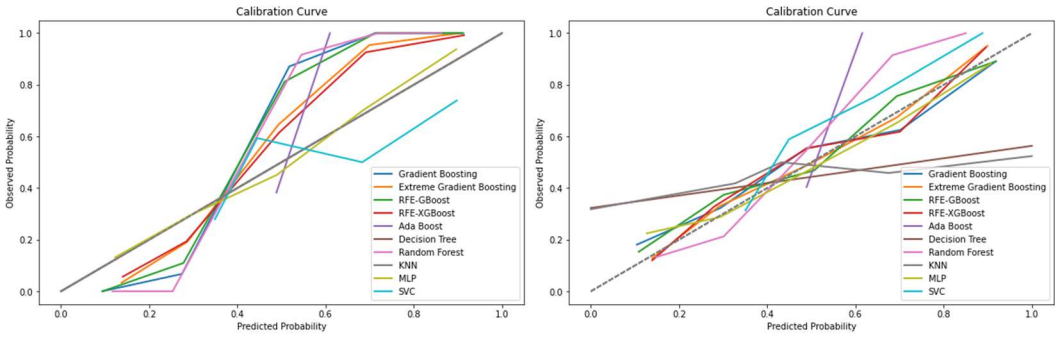

Figure 6. Calibration curve for each model

Calibration curves depicted the calibration in terms of the agreement between the predicted risk of in-hospital mortality and observed in-hospital mortality. The 45° dashed line represents a perfect prediction. The closer the line fit is to the ideal line, the better the predictive accuracy of the model is.

3.8. Classification report

Table 3. Classification report in test set

Model	Class	Precision	Recall	F1-score	Weighted Average F1-score
GB	survive	0.73	0.87	0.79	0.72
	expire	0.74	0.54	0.62	
XGB	survive	0.73	0.87	0.79	0.72
	expire	0.74	0.54	0.62	
RFECV-GB	survive	0.74	0.87	0.8	0.73
	expire	0.75	0.55	0.64	
RFECV-XGB	survive	0.72	0.85	0.78	0.71
	expire	0.71	0.53	0.61	
Ada	survive	0.75	0.83	0.76	0.74
	expire	0.72	0.61	0.6	
DT	survive	0.69	0.75	0.72	0.63
	expire	0.56	0.49	0.52	
RF	survive	0.72	0.9	0.8	0.72
	expire	0.77	0.5	0.61	
KNN	survive	0.63	0.73	0.67	0.56
	expire	0.47	0.34	0.39	
MLP	survive	0.7	0.85	0.77	0.69
	expire	0.69	0.48	0.57	
SVC	survive	0.6	1	0.75	0.46
	expire	1	0.02	0.04	

3.9. Cross validation

Table 4. 5-fold cross validation

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
GB	0.7238	0.7448	0.7573	0.7280	0.6975	0.7303
XGB	0.7280	0.7782	0.7573	0.7406	0.6681	0.7345
RFECV-GB	0.7155	0.7280	0.7573	0.7197	0.7143	0.7270
RFECV-XGB	0.7322	0.7782	0.7322	0.7448	0.6807	0.7336
AdaBoost	0.7238	0.7448	0.6946	0.7113	0.6933	0.7136
DT	0.6067	0.6862	0.6318	0.6192	0.6261	0.6340
RF	0.7657	0.7657	0.7657	0.7406	0.6975	0.7470
KNN	0.6192	0.5983	0.5941	0.6234	0.6092	0.6089
MLP	0.7113	0.6109	0.7197	0.7322	0.6891	0.6926

SVC	0.6192	0.6067	0.6151	0.6025	0.5756	0.6038

The average scores were calculated for the sake of following model selection.

3.10.Confusion matrix

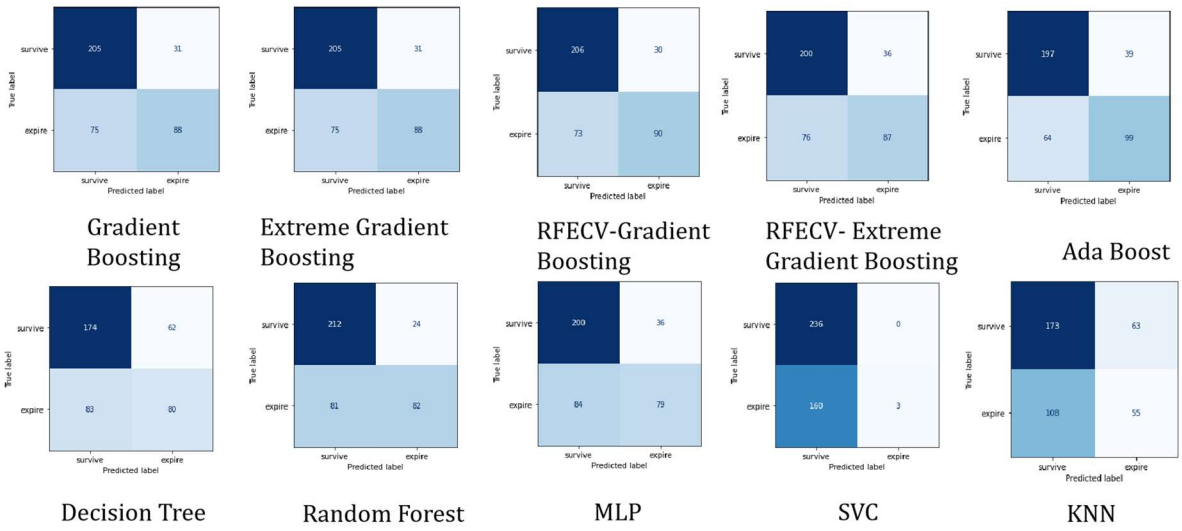

Figure 7. Confusion matrix for each model

Advanced boosting algorithms perform better generally, where as some classifiers like SVC and KNN barely have discrimination capability to distinguish the survival group and expired group.

3.11.Model selection

Table 5. Overall score for each model

Model	AUC	CV	F1-score	Average score
GB	0.8	0.730290776	0.72	0.750096925
XGB	0.8	0.734450266	0.72	0.751483422
RFECV-GB	0.8	0.72695756	0.73	0.752319187
RFECV-XGB	0.78	0.733623992	0.71	0.741207997
Ada	0.79	0.713550858	0.74	0.747850286
DT	0.61	0.633996694	0.63	0.624665565
RF	0.8	0.74702718	0.72	0.755675727
KNN	0.57	0.608878028	0.56	0.579626009
MLP	0.73	0.69262684	0.69	0.704208947
SVC	0.7	0.603828978	0.46	0.587942993

According to the numeric metrics, top 5 models (colored in red) were selected for

feature importance analysis, followed with risk factors identification.

3.12.Permutation feature importance

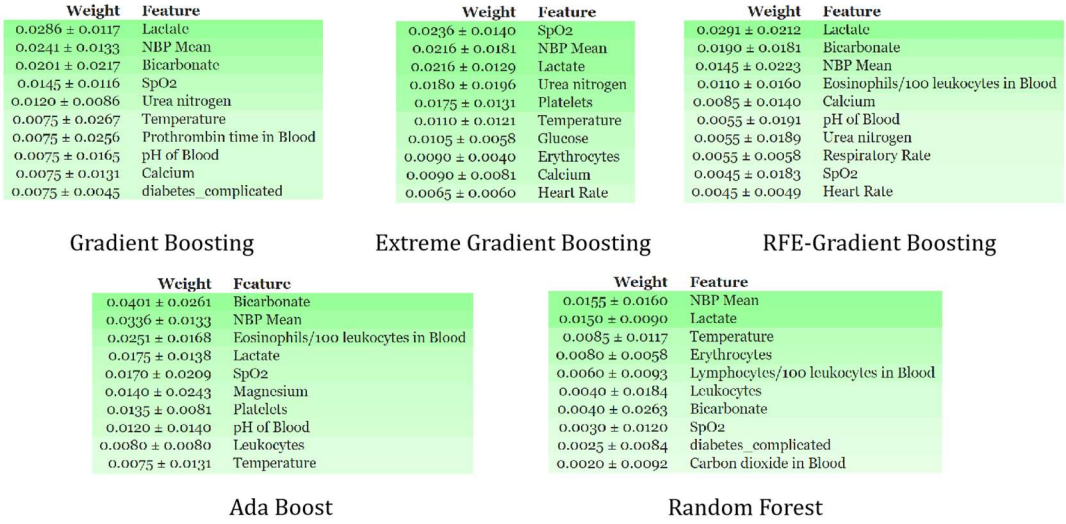

Figure 8. Permutation feature importance of selected models

3.13.Feature selection

Table 6. The occurrence number of each important variable

Variable	Occurrence number	Class
SpO2	5	Vital signs
NBP Mean	5	Vital signs
Lactate	5	Blood examination
Bicarbonate	4	Blood examination
Temperature	4	Vital signs
pH of Blood	3	Blood examination
Calcium	3	Blood examination
Urea nitrogen	3	Blood examination
diabetes complicated	2	Comorbidity
Eosinophils/100 leukocytes in Blood	2	Blood examination
Leukocytes	2	Blood examination
Erythrocytes	2	Blood examination
Heart Rate	2	Vital signs
Platelets	2	Blood examination

3.14.Partial dependence plot

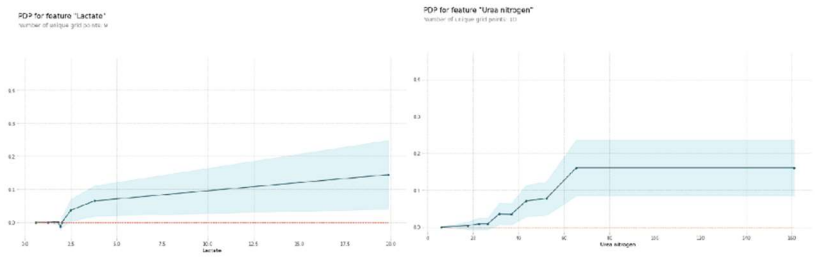

Figure 9. PDP of variables positive correlated to in-hospital mortality

Lactate and urea nitrogen are positive correlated to in-hospital mortality.

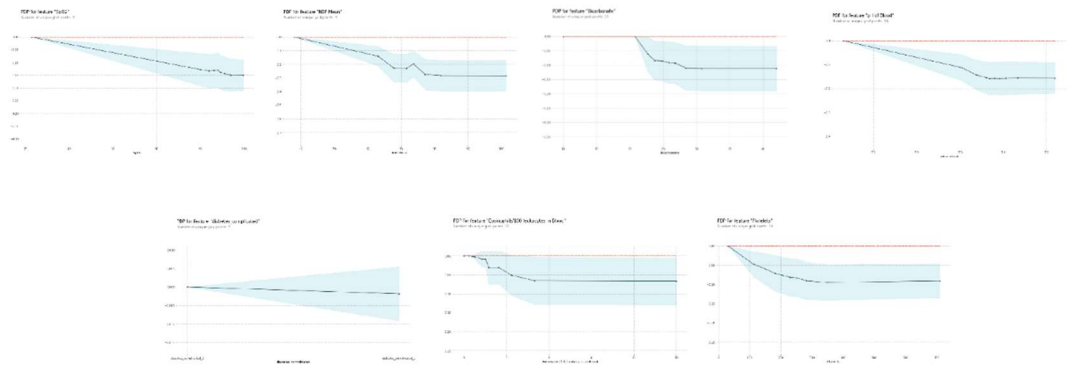

Figure 10. PDP of variables negative correlated to in-hospital mortality

SpO2, NBP Mean, Bicarbonate, pH of Blood, diabetes_complicated, Eosinophils/100 leukocytes in Blood and Platelets are variables negative correlated to in-hospital mortality.

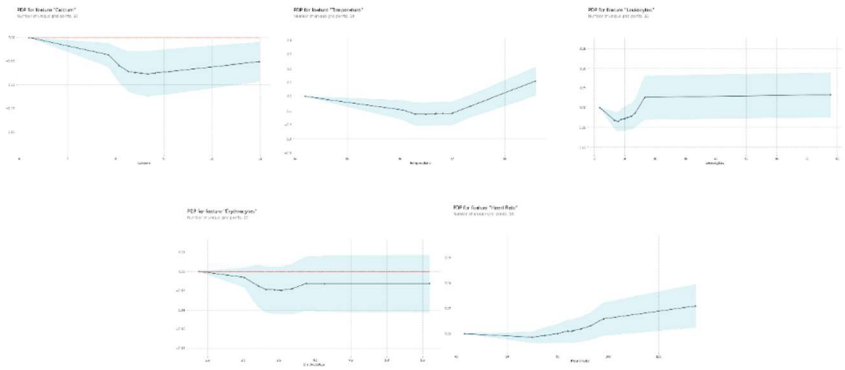

Figure 11. PDP of variables non-linearity correlated to in-hospital mortality

For Calcium, Temperature, heart rate, Leukocytes and Erythrocytes, only appropriate range is safe.

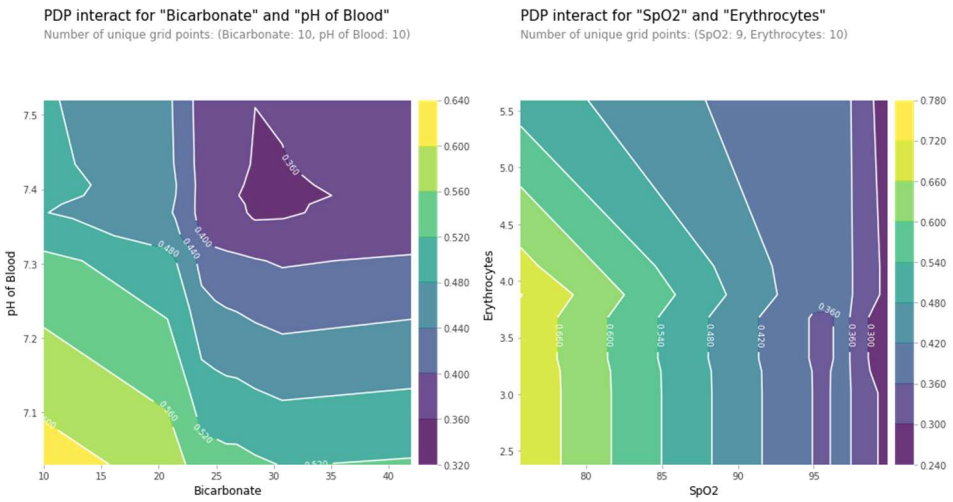

Figure 12. 2-dimension PDP of Bicarbonate and pH of Blood, SpO2 and Erythrocytes

4. Discussion

4.1. Model

Learning curve is used to explore how much the estimator benefits from more data. For XGB, RFECV-XGB, Ada Boost and MLP: training and test scores converge together as more data is added, so, they will probably not benefit from more data.

For other models, the training score is much greater than the test score, which means they probably requires more training examples in order to generalize more effectively.

According to the model evaluation, ensemble learners perform better than others generally. Among the ensemble learners, Extreme Gradient Boosting exhibits the best discrimination capability. RFE strategy didn't improve the performance significantly, which implies that the number of features is appropriate for the task.

4.2. Risk factors

14 variables are identified as risk factors of in-hospital mortality fir heart failure patient, encompassing SpO2, NBP Mean, Lactate, Bicarbonate, Temperature, pH of Blood, Calcium, urea nitrogen, diabetes complicated, Eosinophils/100 leukocytes in Blood, Leukocytes, Erythrocytes, Heart Rate and Platelets. They should be paid close attention during ICU days for heart failure patients. The conclusion may also helpful to clinical decision-making, because it's hard to prioritize the therapy under the complex circumstance in ICU.

10 of them are verified in a similar study[2], which is based on MIMIC III: Lactate, calcium, urea nitrogen, platelet, Leukocytes, Erythrocytes, heart rate, NBP, SpO2 and lactate.

Some insights were obtained from 2-dimension PDP. Bicarbonate, which is an alkaline substance, plays a role as buffer to maintain pH in our body. In solution, the higher the concentration of bicarbonate, the higher the pH. The relationship between bicarbonate and pH of blood with risk are similar. The higher the bicarbonate or pH of blood, the lower the risk. Besides, erythrocytes are cells transporting oxygen in our blood. SpO2 means blood oxygen saturation. Both of them show the negative correlation to the risk. The higher the erythrocytes and the SpO2, the lower the risk.

These evidences verify that the results are reliable and convinced.

5. Future improvement

The vital signs should be measured on only first 24 hours in ICU instead of entire ICU days, because the former could be more representative.

Life support measures should be included in analysis, like mechanical ventilation, renal replacement therapy, etc.

6. Conclusion

According to the clinical data in MIMIC II, the risk factor of in-hospital mortality in heart failure patients were identified, encompassing SpO2, NBP Mean, Lactate, Bicarbonate, Temperature, pH of Blood, Calcium, Urea nitrogen.

They should be paid close attention during ICU days for heart failure patients. The conclusion may also helpful to clinical decision-making, because it's hard to prioritize the therapy under the complex circumstance in ICU.

Generally, blood examination indicators and vital signs play predominant role in contrast with demographical characteristic and comorbidity.

7. Reference

1. MD, A.S., *MIMIC2 original ICU*. 2021.

2. Li, F., et al., *Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database.* BMJ Open, 2021. **11**(7): p. e044779.

3. Zhu, Y., et al., *Machine Learning Prediction Models for Mechanically Ventilated Patients: Analyses of the MIMIC-III Database.* Frontiers in Medicine, 2021. **8**.

4. Hou, N., et al., *Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost.* Journal of Translational Medicine, 2020. **18**(1): p. 462.

5. Breiman, L., *"Random Forests".* Machine Learning, 2001. **45**(1): p. 5-32.

6. Pedregosa., *Scikit-learn: Machine Learning in Python.* JMLR 12, 2011: p. 2825-2830.

7. Lundberg S and L. SI., *A Unified Approach to Interpreting Model Predictions.* NIPS 2017, 2017.

8. Abdul-Rahman, A., W. Morgan, and D.Y. Yu, *A machine learning approach in the non-invasive prediction of intracranial pressure using Modified Photoplethysmography.* PLoS One, 2022. **17**(9): p. e0275417.

9. Zoabi, Y., S. Deri-Rozov, and N. Shomron, *Machine learning-based prediction of COVID-19 diagnosis based on symptoms.* NPJ Digit Med, 2021. **4**(1): p. 3.

10. Fleischhauer, V., A. Feldheiser, and S. Zaunseder, *Beat-to-Beat Blood Pressure Estimation by Photoplethysmography and Its Interpretation.* Sensors (Basel), 2022. **22**(18).