Machine Learning-based Prediction of Postoperative 30-days Mortality

Linna Wang
College of Computer Science, Sichuan University, Chengdu, China
wlnlena@stu.scu.edu.cn

Linji Li
Department of Anesthesiology, West China Hospital, Sichuan University, Chengdu and Department of Anesthesiology, The Second Clinical College of North Sichuan Medical College, Nanchong Central Hospital, Nanchong, China
llj-stephen@163.com

Tao Zhu
Department of Anesthesiology, West China Hospital, Sichuan University
xwtao_zhu@sina.cn

Congli Ma
College of Computer Science, Sichuan University, Chengdu, China
mclzone97@gmail.com

Li Lu*
College of Computer Science, Sichuan University, Chengdu, China
luli@scu.edu.cn

ABSTRACT

Surgical patients aged 65 and over are facing a 2-10 times higher risk of death after surgery. Early prediction of postoperative mortality is essential, as timely and appropriate treatment can improve survival outcomes. With the development of medical and computer technology, numerous available health-related data can be recorded for research. Among various patient indicators which may affect the accuracy of prediction, it is necessary to find highly relevant and efficient features. The aims of this study were to use machine learning algorithms, specifically Bagging and Boosting Algorithms (e.g. Random Forest, eXtreme Gradient Boosting), to predict the postoperative 30-days mortality in surgical patients aged over 65, and to identify the optimal features using genetic algorithm(GA). This prospective study was developed and validated on the cohort from electronic health records (EHRs) of West China Hospital, Sichuan University, which contained 7467 surgical patients (0.924% mortality rate) who underwent surgery between July 1, 2019 and October 31, 2020. Compared with models like the traditional logistic regression model and the baseline ASA physical status, We found that XGBoost with hyper-parameters had best performance based solely on the automatically obtained features (area under the curve [AUC] of 0.9318, 95% confidence interval [CI] 0.9041 - 0.9594). The AUC of baseline ASA-PS was 0.6787 (95% CI 0.6471 - 0.7103) using XGBoost. When both ASA-PS and the selected features are included as inputs, XGboost achieved the AUC of 0.9345 (95% CI 0.9076 - 0.9613).

CCS CONCEPTS

• **Applied computing**; • **Life and medical sciences**; • **Health informatics;**; • **Computing methodologies**; • **Machine learning**; • **Machine learning algorithms**;

KEYWORDS

Imbalanced data, Postoperative mortality, Feature selection, Machine learning, eXtreme gradient boosting, ASA physical status score, Prospective study

ACM Reference Format:
Linna Wang, Linji Li, Tao Zhu, Congli Ma, and Li Lu*. 2021. Machine Learning-based Prediction of Postoperative 30-days Mortality. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487075.3487130

1 INTRODUCTION

An estimated 313 million surgical procedures are undertaken worldwide each year [1]. At least 4.2 million people worldwide die within 30 days of surgery each year [2]. The third greatest death contributor which accounts for 7.7% of all deaths globally is postoperative death [3]. A large proportion of postoperative mortality occurs in a small group of patients with high-risk characteristics, but of this group, less than 15% are admitted to the intensive care unit (ICU) postoperatively [4]. With limited infrustruture, prompt identification of patients on the risks of postoperative mortality and need for critical care monitoring after surgery are vital [5].

With the population ages, the proportion of elderly patients undergoing surgery will continue to rise [6]. As elderly patients got the high number of comorbidity, low tolerance, easy damage of vital organs and poor recovery function, the risks of perioperative severe complications and death in elderly patients are much higher than that in young patients [7]. Many studies have demonstrated that early intervention for comorbidity and surgical conditions in elderly patients can help to reduce or even prevent serious perioperative complications and improve prognosis [5] [8] [9] [10]. In the current

CSAE 2021, October 19–21, 2021, Sanya, China

ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487130

healthcare environment, where health insurance funds are tight and medical personnel is limited, early and rapid identification of elderly patients at high risk of serious perioperative complications and death, and then timely intervention with corresponding strategies, are essential to improve patient outcomes and the allocation of healthcare resources.

To identify people at risk of poor prognosis, most of the existing traditional prediction models (such as logistic regression analysis) and scoring systems are used for risk assessment. There is a lack of a more generalized approach to assessing perioperative risk in older patients.

In recent years, machine learning methods [11] [12] have been gradually applied to the mining and analysis of multi-source heterogeneous medical big data from electronic medical records in order to improve the performance of various models. There are previous researches showed that machine learning may offer better predictive performance when data input are abundant and variable interactions are complex [13]. A vital key is to have approches to rapidly identify patients who are at high risk and most in need of manual intervention. The prediction models can help to provide strong basis for clinicians making decision and reasonable allocation of public medical healthcare resources.

In this study, we implemented machine learning methods to accurately predict postoperative mortality in elderly patients using available information from EHRs and achieved quantitative assessment of perioperative risk. Genetic Algorithm [14] was used to select available optimal features using data accumulated from a structured data platform of West China Hospital, Sichuan University.

2 RELATED WORKS

Many patients will choose surgery to cure a disease or to prolong their lives, a significant growth in the demand for surgical services will be result from the aging of the population [15]. Related studies suggest short-term postoperative mortality varies from 1-3%, while high risk surgery group accounts for 80% of the deaths [16] [17]. Postoperative mortality prediction is now a much researched filed. There are several existing preoperative patient risk scores have been developed for this purpose, such as the American Society of Anesthesiologists physical status (ASA-PS), Preoperative Score to Predict Postoperative Mortality (POSPOM), Surgical Outcome Risk Tool (SORT), Charlson Comorbidity Index, and American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) [18] [19] [20] [21]. In order to create models of risk, some of these tools need to leverage International Statistical Classification of Diseases and Related Health Problems (ICD) codes which means that data used for analysis are not available prior to the procedure. Previous studies discussed about their lack of precision at the patient level. Scores like ASA score can not be fully automated because it relies on the subjective judgement of clinicians who are well trained [22] [23].

The most commonly used structured medical measurements can be classified as numerical type and categorical type. With these available indicators, many mortality prediction models based on machine learning algorithms or deep learning algorithms were proposed to evaluate patients risk of death. Machine learning (ML) methods have bee proven to have potential in clinical risk prediction. Pirracchio et al. used an ensemble ML technique which utilized multiple algorithms to obtain better performance compared with SAPS-II, APACHE-II, and SOFA [24]. Their approach showed an AUC of around 0.88 which compared to AUC of 0.78 generated by SAPS-II. To predict postoperative in-hospital mortality, Hill B L et al. reported on the use of ML algorithms to create a fully automated score [23]. They found random forest classifier outperformed other traditional scores while using obtained features.

It is also important to select the best subset form abundant structured features. Filter methods, wrapper methods and embedded methods are three main approaches to select features [25] [26]. Babatunde et al. detailed the application of a binary Genetic Algorithm(GA) for dimensional reduction to enhance the performance of classifiers [14].

These related works show that ML and GA techniques have potential to predict postoperative mortality. However, patients in these studies were specified with specific conditions or not from Asian datasets. Therefore, the results can not respond well to Asian elderly patients. As clinicians' consensus account for a greater proportion of feature selection, it is not conducive to breaking out of the old framework. In summary, we proposed a mortality prediction method which only using easy-to-extract perioperative electronic health record data and is more broadly utilizable at Asian hospital.

3 METHODS

3.1 Dataset

We presented the data source and study population in this section. A single-center cohort prospective analysis was conducted, consisting of 57,728 patients who underwent surgery between July 1, 2019 and October 31, 2020 from EHRs of West China Hospital, Sichuan University. In our study, only patients aged 65 years or above were included to predict the risk of mortality. The high-quality records with sufficient amount of data measurements such as patients' demographics, comorbidity, operative characteristics, and preoperative laboratory tests which can influence the outcome were extracted. Preoperative comorbidity were recorded using International Statistical Classification of Disease and Related Health Problems, 10th Revision (ICD-10) codes [27]. Briefly, in the first stage, we extracted data into 4 main tables structured around 4 distinct concepts: patients, preoperative data, intraoperative events and postoperative events. Data in these tables are then used to populate a series of measures and metrics such as ASA physical status, glucose, and others. All patients' data were anonymized prior to extraction and analysis. This study was approved by the ethics review board of West China Hospital, Sichuan University.

3.2 Model Endpoint Definition

We trained classification models to predict the postoperative mortality within 30 days as a binary outcome. The death flag of each patient was recorded in the postoperative events table. We used the date of postoperative event minus the date of the end of surgery to obtain how many days after the surgery. This classification was set to true if, within 30 days, death flag was recorded as '1'. Because of

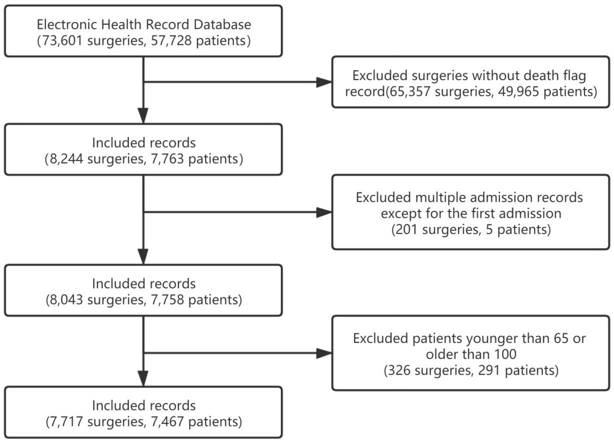
Electronic Health Record Database
(73,601 surgeries, 57,728 patients)
Excluded surgeries without death flag record(65,357 surgeries, 49,965 patients)
Included records
(8,244 surgeries, 7,763 patients)
Excluded multiple admission records except for the first admission
(201 surgeries, 5 patients)
Included records
(8,043 surgeries, 7,758 patients)
Excluded patients younger than 65 or older than 100
(326 surgeries, 291 patients)
Included records
(7,717 surgeries, 7,467 patients)

Figure 1: The Detailed Process of Data Extraction.

the the possibility of false positive samples, we had clinicians from Huaxi hospital to validate one subset of samples.

3.3 Inclusion and Exclusion Criteria

At the unique patient level, 57,728 patients who underwent a surgery between July 1, 2019 and October 31, 2020 were included in the study. Cases on patients older than 65 yr of age and younger than 100 yr of age were included, which contains 13,803 unique patients. Because a patient may have multiple admission records, only the first admission was analyzed. Some patients underwent more than one procedure during their hospitalization, a phenomenon that occured more frequently in the high-risk patient population. The final study included all surgeries which had postoperative events records and met the above criteria, which contains 7,467 unique patients and 7,717 unique surgeries. We list the process of the inclusion and exclusion for overall data set in Figure 1

We performed an analysis to verify that including these surgeries did not unduly affect the distribution of correlations between pairs of surgeries compared to only including the first surgery of each admission. For our data set and first-surgery-only data set, we randomly selected 100,000 pairs of surgeries from two different patients in each data set, and then calculated the Pearson correlation coefficients [28], using the features including AGE, GLUCOSE, WHITE BLOOD CELL COUNT, PULSE, WEIGHT, BMI, BILIRUBIN TOTAL, CREATININE, POTASSIUM, ALBUMIN, HEIGHT, SPO2, GENDER, ASA_STATUS, CARDIAC FUNCTION, MOVEMENT EQUIVALENT, HYPERTENSION, and SMOKE [23]. The distribution of correlation coefficients was shown in Figure 2 which could indicate that the distributions were similar. Note that, to against information leaks, patients appear in testing set were removed from training set.

3.4 Data Preprocessing

Each surgical record corresponded to a unique hospital admission ID. As it is common with missing data in the records, before feature selection, we removed the variables with more than 30% missing to facilitate and ensure the accuracy of the research. For numeric variables with less than 5% missing data or randomly missing data, we filled the missing values with the median value for the respective feature. For variables with a missing rate between 5%-30%, SoftImpute algorithm [29] was carried out to fix the problem leveraging the similarity of groups of patients. If the observation value was clinically beyond the top and bottom 1% of the actual distribution, we set these outliers to random numbers from 1% to 25% percentiles and 75% to 99% percentiles, respectively. In addition, we standardized the numerical data. After the data is centralized according to the mean value, and then scaled according to the standard deviation, the data followed the normal distribution with a mean of 0 and a variance of 1. For categorical features, missing values were treated as a new category. Categorical features were converted into numerical values by one-hot encoding scheme.

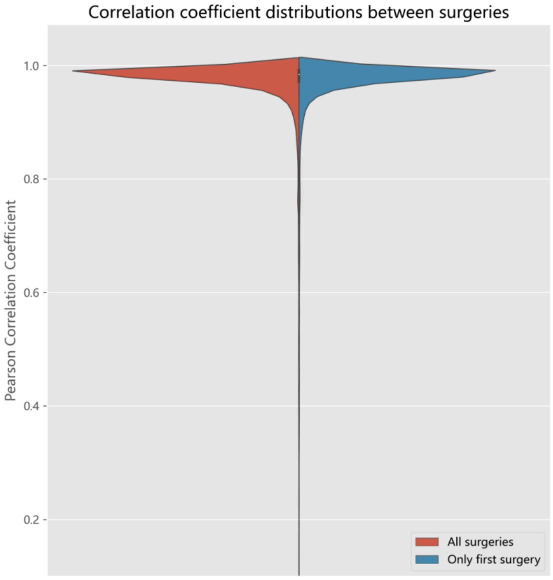

Figure 2: Correlations between Pairs of Surgeries in the Same Admission and Randomly Selected Pairs of Surgeries. Violin Plot Comparing the Distribution of Pearson Correlation Coefficients for 100,000 Randomly Sampled Pairs of Surgeries from the Set of All Surgeries (Left, in Red) and from the Set of only the First Surgery in Each Admission (Right, in Blue). The Distributions Were Identical.

To make sure there were similar mortality rates for training and testing, we extracted death samples and survival samples separately, then randomly divided both sample sets into a large set (70%) and a small set (30%). Finally, we merged the two large sample sets as training set (70%, n = 5406) and merged the two small sample sets as testing set (30%, n = 2311) while the same patient did not appear

Table 1: Patient Characteristics Employed for Training and Testing Models. 'n' Denots the Number of Patients

Property	Training data	Testing data
Patient, n	5226	2241
Admission, n	5319	2285
Surgery, n	5406	2311
Average surgery numbers of one patient	1.034	1.031
Average admission numbers of one patient	1.018	1.02
Average surgery numbers of one admission	1.016	1.011
Patients with not just one admission, n (%)	180(3.44)	70(3.12)
Admissions with not just one surgery, n (%)	88(1.65)	26(1.14)
Mortalities, n (%)	45(0.83)	24(1.04%)
Mean age	72.1	71.8
Male patients, n (%)	3103(57.40)	1268(54.87)
ASA physical status 1, n (%)	11(0.20)	10(0.43)
ASA physical status 2, n (%)	2603(48.15)	1126(48.72)
ASA physical status 3, n (%)	2733(50.55)	1159(50.15)
ASA physical status 4, n (%)	54(0.99)	16(0.69)
ASA physical status 5, n (%)	5(0.09)	0(0)
Types of surgery, n (%)		
Abdominal surgery	2992(55.35)	1276(55.21)
Orthopedic surgery	1003(18.55)	433(18.74)
Thoracic surgery	503(9.30)	188(8.14)
Cardiac surgery	251(4.64)	90(3.89)
Craniocerebral surgery	6(0.11)	5(0.22)
Other	651(12.04)	319(13.80)

in the same set to protect information from leaking. Table 1 was made for the detailed patient characteristics information.

To solve the extreme imbalanced classification between the death and survival(0.89% mortality rate), we used the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [30] on training set to synthesize new examples from the minority class. The SMOTE implementation is provided by the imbalanced-learn Python library in the SMOTE class [31]. We set a more balanced distribution with 500 dead samples and 5,361 surviving samples. To maintain the natural outcomes, our testing set was not upsampled.

3.5 Model Input Features

Preoperative laboratory tests were the latest taken before the start time of surgery. Surgery details included anesthesia type, surgery type, duration of operation, ASA-PS class(under 6), vital signs(such as systolic blood pressure (SBP), diastolic blood pressure (DBP) and pulse), and others.

After data cleaning, we first got 177 initial available variables to start feature selection. The subset of features obtained from feature selection should be as small and effective as possible, in the meantime, improve but not reduce the accuracy of prediction. There were two main purposes for feature selection in this study: 1) to reduce model over-fitting and improve model generalization ability by feature reduction; 2) to enhance the understanding between features and feature values [32] [33]. The search strategies for feature selection are classified as: complete search strategy (e.g. Breadth First Search), heuristic search strategy (e.g. Sequential Forward Selection) and stochastic search algorithm (e.g. Genetic Algorithm [34] [35] [36]). To decide how many and which features we should put into training, Genetic Algorithm (GA) was implemented to determine the optimal features subset while using Decision Tree [37] as estimator. The main operators of the genetic algorithms are reproduction, crossover, and mutation. We used DEAP to implement GA [38]. For DEAP global variables of GA, we set 1) tournament selection as selection operator with tournament size of 3, 2) Single Point Crossover as crossover, 3) multiple Flip-bit mutation as mutation operator with indpb (independent probability for each attribute to be flipped) of 0.05. Population size was set to 100. Recursive Feature Elimination(RFE) [39] was used to validate the selected important features. The number of optimal features was 81 with automatic tuning of the number of features selected with 5-fold cross-validation. During the feature selection, we plotted (Figure 3) the test and validation classification accuracy to see how these numbers changed as we getting close to the best feature subsets.

Combined with West China Hospital clinicians' consensus, we finally had 73 features (including ASA status) to be potentially predictive of the postoperative mortality within 30 days. The full list of features included basic demographic information such as age, gender, and Body Mass Index; available obtained laboratory tests prior to surgery, for example albumin and total bilirubin; descriptive intraoperative vital signs, such as diastolic pressure values; summary of drug and fluid interventions, such as total fluid transfused; comorbidity, such as diabetes; and patient surgery descriptions, such as surgery type. This study included the most recent values of the variables prior to the surgery.

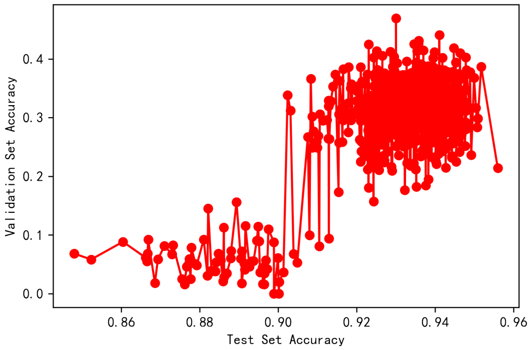

Figure 3: Test and Validation Classification Accuracy.

To facilitate the analysis of the impact of different feature inputs on prediction, we built 3 types of input. Type 1 included all the 73 features. Type 2 removed the ASA physical status score while included other 72 features - as ASA score cannot be fully automated until they are reviewed by a trained anesthesiologist. The baseline models were also tested: Type 3 included only the ASA physical status score. The details of all types can be found below.

3.6 Model Creation, Training, and Testing

This study considered 4 different models due to their widespread use: Random Forests (RF), Logistic Regression (LR), XGBoost, and Multi-Layer Perceptron (MLP). From Bagging and Boosting ensemble learning algorithms, we chose 2 algorithms to develop this study:

- RF [40] is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the datasets and uses averaging to improve the predictive accuracy and control over-fitting.
- XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable [41]. XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way. It is a popular method used in medical field.

A benefit of using the above algorithms is that after trees are constructed, it is relatively straightforward to retrieve importance scores for each attribute. They have the property of 'feature importance' which is handy to get impurity-based feature importance. Generally, we can find out how useful and valuable each feature was in the construction of model. The 73 features and feature importance are shown and described in Supplementary Appendix. In addition, common linear and nonlinear classifiers were also included:

- (3) Logical regression algorithm, which is a classic binary classification algorithm, has good robustness to small and medium-sized noises of data after the mapping of Sigmoid function.
- (4) MLP, a feed-forward artificial neural network, can distinguish non-linearly distinguishable data [42]. An MLP is characterized by several layers of input nodes connected as a directed graph between the input nodes connected as a directed graph between the input and output layers.

Model hyper-parameters were chosen by using RandomizedSearchCV and GridSearchCV with five-fold cross validation on the training set. One unique patient should not appear in both training and testing fold, but only one fold. All 4 classifiers were set with default parameters before parameter optimization. First, we set a large range of values for the parameters in each model and used RandomizedSearchCV with 200 iterations to search the best parameters. Next, a smaller range was determined based on the parameter selected in the previous step, and then GridSearchCV was used to fine tune the parameters. Since we used five-fold cross-validation which meant that there were five divided data partitions for training and testing. In this case, every one of the five partitions was treated as a testing set only once and a training set four times. Through multiple trials, average metrics could be obtained which helped to get a better assessment of model performance.

We imported RandomForestClassifier, LogisticRegression and MLPClassifier from scikit-learn [43]. The XGBClassifier was implemented using the XGBoost package [41]. All performance metrics were implemented by using Scikit-Learn.

Block bootstrapping was used to help generate confidence intervals (CIs) for the performance metrics on test set. Prediction correlation appears in surgeries from the same patient. Nevertheless, in order to avoid the lose of the correlation structure, instead of sampling surgeries randomly, we sampled patients randomly, and included all predictions in the bootstrap sample. We set this procedure to repeat 1,000 times and calculated performance metrics of each bootstrap sample. After these metrics got sorted, the 95% CI was determined by selecting the 25th and 975th values of the sorted metrics.

3.7 Evaluation Metrics

Receiver operating characteristic curves (ROC) is also know as sensitivity curve. Each time a different threshold is selected so that a set of FPR and TPR is obtained, with the FPR value as the horizontal coordinate and the TPR value as the vertical coordinate, we got ROC curve. It illustrates how the model performance varies with the threshold value. Model performance was first assessed using area under the ROC of 95% CI which was calculated using bootstrapping with 1,000 samples. As we were dealing with imbalanced data, accuracy should not be the only metric to dedicate model performance. We also calculated Precision, Recall and F1 to evaluate the performance of classifiers in a much better way. True Positive (TP) is the number of truly classify as a positive, and False Positive (FP) is the number of truly classify as a negative. False Negative (FN) is the number of falsely classified as negative. True Negative (TN) is the number of falsely classified as positive. Precision is the accuracy of the positive predictions. Recall is the ratio of positive instances that are correctly detected. Combine precision and recall into a single metric which is F1 score. We used sklearn. metrics function to compute these scores. In the mean time, the precision-recall curve was plotted and the performance of a model was indicated by observing whether the curve reached the point in the upper right corner. In this paper, all of this was implemented in Python.

Table 2: Area under the Receiver Operating Characteristic (AUROC) Curve. The Highest AUROC Are Shown in Bold. The Mean Value of the AUROC Is Shown, along with the 95% Confidence Interval (CI) in Parenthesis

Model/AUC (95% CI)	Random Forests	XGBoost	Logistic regression	MLP
ASA status	0.6782 (0.646 - 0.7104)	**0.6787 (0.6471 - 0.7103)**	0.6759 (0.6431 - 0.7087)	0.6773 (0.6447 - 0.7098)
Selected features	**0.9955 (0.9937 - 0.9973)**	0.9318 (0.9041 - 0.9594)	0.8495 (0.8153 - 0.8836)	0.8725 (0.8272 - 0.9179)
Selected + ASA status	**0.9949 (0.9925 - 0.9972)**	0.9345 (0.9076 - 0.9613)	0.8514 (0.8181 - 0.8847)	0.8717 (0.8246 - 0.9188)

Table 3: Performance Metrics for All Models Using Selected Features

Selected features (95% CI)	Random Forests	XGBoost	Logistic regression	MLP
Accuracy	0.9917 (0.9903 - 0.9932)	**0.9937 (0.9922 - 0.9951)**	0.98 (0.9766 - 0.9835)	0.9808 (0.9771 - 0.9844)
Precision	**0.9865 (0.973 - 1.0)**	0.9773 (0.9545 - 1.0)	0.6968 (0.5869 - 0.8066)	0.6989 (0.5854 - 0.8125)
Recall	0.6909 (0.6364 - 0.7455)	**0.7727 (0.7273 - 0.8182)**	0.473 (0.4182 - 0.5277)	0.5182 (0.4364 - 0.6)
F1 score	0.816 (0.7778 - 0.8542)	**0.8667 (0.8333 - 0.9)**	0.5653 (0.5055 - 0.625)	0.5852 (0.5106 - 0.6598)

4 RESULTS

4.1 Patient Characteristics

We contained 7,717 surgical records encompassing 7,467 patients in this study. Patients were between the ages of 65 and 100 yr, with a mean age of 72 yr. The postoperative 30-days mortality was approximately 0.924%. ASA score of 3 was the most common, comprising 50.5% of the data set. In Table 1, we provided further information on patient characteristics.

4.2 Model Performance

To evaluate our models, we plotted 2 curves: ROC and precision-recall curve. We bolded the best scores of each metrics.

Area under the ROC

The area under the ROC curve values of each model with different input features were recorded in Table 2. The ROC curve of different models using selected features and ASA as input was shown in Figure 4

Precision-recall

The precision-recall (PR) curve using the XGboost for the three feature sets is shown in Figure 5. The PR curve of all models on all-73-feature set are shown in Figure 6. Generally, the selected feature set performed better than other two.

Calibration

Brier score is another evaluation metric used in this paper, which calculated the mean squared error between predicted probabilities and the actual values. The value of Brier score is always between 0.0 and 1.0, where a model with better performance has a lower brier score. The implementation of brier score was using sckit-learn's inbuilt function. By observing the Brier scores it can be found that the nonlinear models performed better compared to the linear models. When using ASA status as the only input, the lowest Brier score (0.098) was achieved by the random forest and XGBoost classifiers. When using the selected features with ASA status as input, the XGBoost classifier had the lowest Brier score of 0.0043.

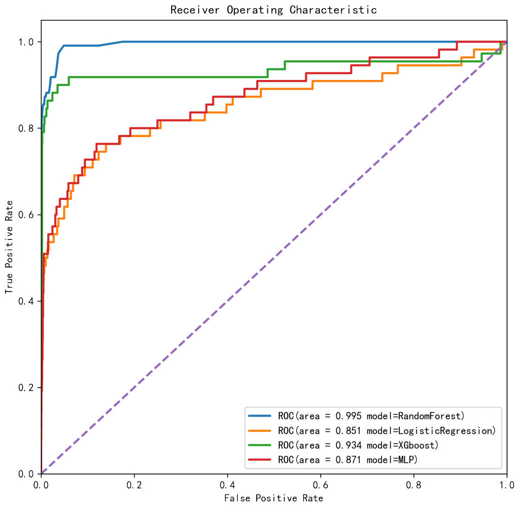

Figure 4: ROC Curve of Different Models Using Selected Features and ASA-PS. The Non-Linear Models Outperformed the Linear Models. In Particular, the Random Forest Had the Highest AUROC Compared with the Other Models.

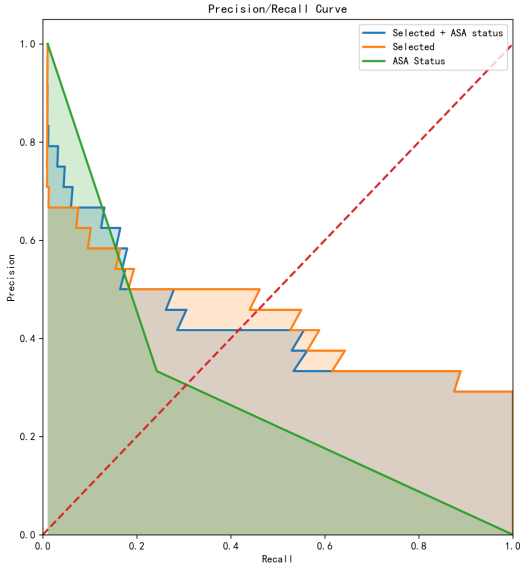

Figure 5: PR Curve Using the XGboost for the Three Feature Sets.

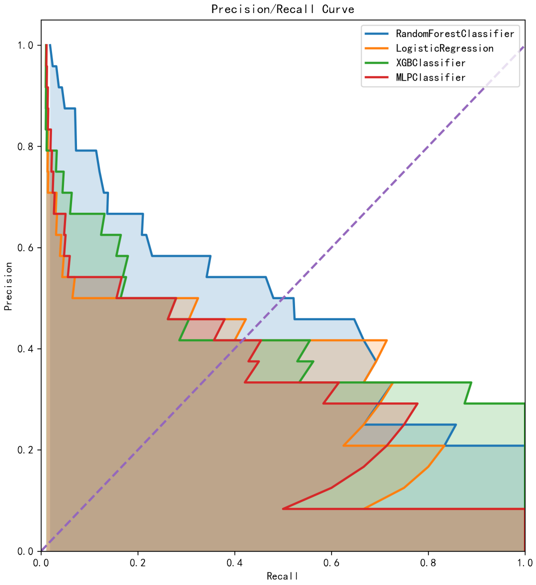

Figure 6: PR curve of All Models Using Selected Features and ASA-PS.

5 DISCUSSION

Table 3 and Table 4 contained the Accuracy, Precision, Recall and F1 for all 4 models. Combined with all these metrics we got, we found that XGboost outperformed other models. Confidence intervals derived by bootstrapping were shown in parenthesis.

Models using the selected features had higher area under the ROC values (XGboost 0.9318, 95% CI 0.9041 - 0.9594) than baseline ASA-PS (XGboost 0.6787, 95% CI 0.6471 - 0.7103). Table 2 showed that adding ASA-PS values to the selected features did not much improve performance as compared with the selected features alone. Including the ASA-PS with the selected features achieved an AUC of 0.9345 (95% CI 0.9076 - 0.9613).

The recall metric showed in Table 3 and Table 4 did not reach a high score as Precision did. Here we made a brief description about the 3 reasons may caused this unsatisfactory results. 1) The number of negative data points(majority class) in our datasets were large compared to that of the positive data (minority class). We used SMOTE method which can oversample the minority class to address the question of imbalanced classification. While experiments were done in this study to show that using SMOTE did get higher precision and recall than not using it, the shortcomings of SMOTE still cannot be ignored. SMOTE always assigns a global neighborhood parameter K but neglects the local distribution characteristics, thus resulting in a greater chance of class mixture which affects classification results. 2) The non-survivors sample size was too small(0.89% mortality rate) which increased difficulty for machine learning to learn. Fortunately, the database is still incorporating new cases and a larger amount of data will be available in future experiments. 3) The threshold setting of the predicted probability is also important. The approach of setting weights needs to be improved in future experiments.

Our analysis of this data furnished us with two key insights about prediction of mortality. First, all of the bagging and boosting models achieved high performance. No matter which specific model or feature set(general demographic data, surgical data, and basic lab data) were chosen, all of the models achieved area under the curve values in the 0.67 to 0.99 range based on an extreme imbalanced dataset. The scores (AUC, Accuracy, Precision, Recall, F1) demonstrated the benefit of using a boosting model, XGboost - as opposed to the traditional analysis and ASA score for early prediction of probability of mortality. LR performed worse than other classifiers. Second, we found that there was virtually no loss in performance if our models were restricted to available data even without ASA-PS. These two keys suggest that our models could be used to identify patients that are at high risk of death while still in the hospital and soon after the surgical procedure.

6 CONCLUSIONS

In this paper, all data used for this study were obtained from the custom built perioperative database, which containing all patients who have undergone surgery at West China Hospital, Sichuan University, since July 1, 2019. And the database is still growing. Based on these available data, we explored using Genetic Algorithm for features selection and developed machine learning methods to predict postoperative 30-days mortality in surgical patients aged over 65. We successfully selected a set of features that could be used

Table 4: Performance Metrics for All Models Using Selected Features and ASA-PS

Selected+ ASA (95% CI)	Random Forests	XGBoost	Logistic regression	MLP
Accuracy	0.9927 (0.9912 - 0.9942)	**0.9937 (0.9922 - 0.9951)**	0.9788 (0.9752 - 0.9825)	0.9805 (0.9766 - 0.9844)
Precision	**0.99(0.98 - 1.0)**	0.9878 (0.9756 - 1.0)	0.6584 (0.5455 - 0.7714)	0.6964 (0.5769 - 0.8158)
Recall	0.7273 (0.6727 - 0.7818)	**0.7727 (0.7273 - 0.8182)**	0.4636 (0.4 - 0.5273)	0.5 (0.4182 - 0.5818)
F1 score	0.8409 (0.8043 - 0.8776)	**0.8667 (0.8333 - 0.9)**	0.5412 (0.4783 - 0.6042)	0.577 (0.5 - 0.654)

to better predict in-hospital mortality. This group of features do not necessitate the assistance of a clinician for score calculation in contrast to the ASA physical status. Our selected features achieved a good performance that was comparable to ASA-PS in that model. By using selected available variables, the boosting models (XGboost) was more accurate than other machine learning models. Our study is a strong evidence to illustrate that machine learning models could improve discrimination of the prediction model and identify high risk patients, which can help guide the rational allocation of public health care resources. The promise of using machine learning technology in healthcare is huge. As the MLP method performed not so well in this study, there is still a lot of room for improvement in deep learning methods. Our hope and expectation will be improving models and producing deep researches leveraging unstructured data obtained from surgery in the next few years.

ACKNOWLEDGMENTS

This research was financially supported by Sichuan Province Key Research and Development Plan. Project 2019YFG0491 and Project 2021YFG0304. And National Key R&D Program of China (2018YFC2001800).

REFERENCES

[1] Meara, J.G., *et al.* (2015). Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. The Lancet, 386(9993): p. 569-624.

[2] Nepogodiev, D., *et al.* (2019). Global burden of postoperative death. The Lancet, 93(10170).

[3] GBD 2016 Causes of Death Collaborators (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017; 390: 1151–210.

[4] Pearse RM, Harrison DA, James P, Watson D, Hinds C, Rhodes A, Grounds RM, Bennett ED (2006). Identification and characterisation of the high-risk surgical population in the United Kingdom. Crit Care. 10(3):R81. doi: 10.1186/cc4928. Epub 2006 Jun 2. PMID: 16749940; PMCID: PMC1550954.

[5] Chiew C J, Liu N, Wong T H, *et al.* (2020). Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission[J]. Annals of surgery, 272(6): 1133.

[6] Revenig, L.M., Ogan, K., Guzzo, T.J. *et al.* (2014). The Use of Frailty as a Surgical Risk Assessment Tool in Elderly Patients. Curr Geri Rep 3, 1-7.

[7] Gilbert T, Neuburger J, Kraindler J, *et al.* (2018). Development and validation of a Hospital Frailty Risk Score focusing on older people in acute care settings using electronic hospital records: an observational study[J]. The Lancet, 391(10132): 1775-1782.

[8] Preeya K. Mistry, Geoffrey S. Gaunay, David M. Hoenig (2017). Prediction of surgical complications in the elderly: Can we improve outcomes?,Asian Journal of Urology, Volume 4, Issue 1, Pages 44-49,ISSN 2214-3882.

[9] Aggarwal, G.; Broughton, K.J.; Williams, L.J.; Peden, C.J.; Quiney, N (2020). Early Postoperative Death in Patients Undergoing Emergency High-Risk Surgery: Towards a Better Understanding of Patients for Whom Surgery May not Be Beneficial. J. Clin. Med. 9, 1288.

[10] Naito T, Mitsunaga S, Miura S, *et al.* (2019). Feasibility of early multimodal interventions for elderly patients with advanced pancreatic and non-small-cell lung cancer[J]. Journal of cachexia, sarcopenia and muscle, 2019, 10(1): 73-83.

[11] Han J, Kamber M, Pei J (2011). Data mining concepts and techniques third edition[J]. The Morgan Kaufmann Series in Data Management Systems, 5(4): 83-124.

[12] Witten I H, Frank E (2002). Data mining: practical machine learning tools and techniques with Java implementations[J]. Acm Sigmod Record, 31(1): 76-77.

[13] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[14] Babatunde O H, Armstrong L, Leng J, *et al.* (2014). A genetic algorithm-based feature selection[J].

[15] Etzioni D A, Liu J H, Maggard M A, *et al.* (2003). The aging population and its impact on the surgery workforce[J]. Annals of surgery, 238(2): 170.

[16] The International Surgical Outcomes Study group, Global patient outcomes after elective surgery: prospective cohort study in 27 low-, middle- and high-income countries, BJA: British Journal of Anaesthesia, Volume 117, Issue 5, November 2016, Pages 601–609, https://doi.org/10.1093/bja/aew316.

[17] Glance L G, Lustik S J, Hannan E L, *et al.* (2012). The Surgical Mortality Probability Model: derivation and validation of a simple risk prediction rule for noncardiac surgery[J]. Annals of surgery, 255(4): 696-702.

[18] Le Manach Y, Collins G, Rodseth R, *et al.* (2016). Preoperative score to predict postoperative mortality (POSPOM). Anesthesiology, 124: 570e9.

[19] Protopapa K L, Simpson J C, Smith N C E, *et al.* (2014). Development and validation of the surgical outcome risk tool (SORT)[J]. The British journal of surgery, 101(13): 1774.

[20] Charlson M, Szatrowski T P, Peterson J, *et al.* (1994). Validation of a combined comorbidity index[J]. Journal of clinical epidemiology, 47(11): 1245-1251.

[21] Bilimoria K Y, Liu Y, Paruch J L, *et al.* (2013). Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons[J]. Journal of the American College of Surgeons, 217(5): 833-842. e3.

[22] Cohen M E, Bilimoria K Y, Ko C Y, *et al.* (2009). Effect of subjective preoperative variables on risk-adjusted assessment of hospital morbidity and mortality[J]. Annals of surgery, 249(4): 682-689.

[23] Hill B L, Brown R, Gabel E, *et al.* (2019). An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data[J]. British journal of anaesthesia, 123(6): 877-886.

[24] Pirracchio R, Petersen M L, Carone M, *et al.* (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study[J]. The Lancet Respiratory Medicine, 3(1): 42-52.

[25] Guyon I, Elisseeff A (2003). An introduction to variable and feature selection[J]. Journal of machine learning research, 3(Mar): 1157-1182.

[26] Mao Y, Yang Y (2019). A wrapper feature subset selection method based on randomized search and multilayer structure[J]. BioMed research international, 2019.

[27] WHO. International Statistical Classification of Diseases and Related Health Problems-10th Revision[M]. 1993.

[28] Haldun Akoglu (2018). User's guide to correlation coefficients,Turkish Journal of Emergency Medicine,Volume 18, Issue 3, Pages 91-93, ISSN 2452-2473.

[29] Mazumder R, Hastie T, Tibshirani R (2010). Spectral regularization algorithms for learning large incomplete matrices[J]. The Journal of Machine Learning Research, 11: 2287-2322.

[30] Chawla N V, Bowyer K W, Hall L O, *et al.* (2002). SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 16: 321-357.

[31] Lemaître G, Nogueira F, Aridas C K (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning[J]. The Journal of Machine Learning Research, 18(1): 559-563.

[32] B. Xue, M. Zhang, W. N. Browne and X. Yao (2016). A Survey on Evolutionary Computation Approaches to Feature Selection," in IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, pp. 606-626, doi: 10.1109/TEVC.2015.2504420.

[33] Chandrashekar G, Sahin F (2014). A survey on feature selection methods[J]. Computers & Electrical Engineering, 40(1): 16-28.

[34] Golberg D E (1989). Genetic algorithms in search, optimization, and machine learning[J]. Addion wesley, 1989(102): 36.

[35] Yang J, Honavar V (1998). Feature subset selection using a genetic algorithm[M]//Feature extraction, construction and selection. Springer, Boston, MA,

117-136.

[36] Anbarasi M, Anupriya E, Iyengar N (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm[J]. International Journal of Engineering Science and Technology, 2(10): 5370-5376.

[37] Quinlan J R (1986). Induction of decision trees[J]. Machine learning, 1(1): 81-106.

[38] Fortin F A, De Rainville F M, Gardner M A G, *et al.* (2012). DEAP: Evolutionary algorithms made easy[J]. The Journal of Machine Learning Research, 13(1): 2171-2175.

[39] Guyon I, Weston J, Barnhill S, *et al.* (2002). Gene selection for cancer classification using support vector machines[J]. Machine learning, 2002, 46(1): 389-422.

[40] Breiman L (2001). Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[41] Chen T, Guestrin C (2016). Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794.

[42] Hastie T, Tibshirani R, Friedman J (2009). The elements of statistical learning: data mining, inference, and prediction[M]. Springer Science & Business Media.

[43] Pedregosa F, Varoquaux G, Gramfort A, *et al.* (2011). Scikit-learn: Machine learning in Python[J]. the Journal of machine Learning research, 12: 2825-2830.