BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Explainable predictions of a machine learning model to forecast the postoperative length of stay for severe patients: machine learning model development and evaluation

Ha Na Cho[1], Imjin Ahn[2], Hansle Gwon[2], Hee Jun Kang[1], Yunha Kim[2], Hyeram Seo[2], Heejung Choi[2], Minkyoung Kim[2], Jiye Han[2], Gaeun Kee[1], Seohyun Park[1], Tae Joon Jun[3*†] and Young-Hak Kim[1†]

Abstract

Background Predicting the length of stay in advance will not only benefit the hospitals both clinically and financially but enable healthcare providers to better decision-making for improved quality of care. More importantly, understanding the length of stay of severe patients who require general anesthesia is key to enhancing health outcomes.

Objective Here, we aim to discover how machine learning can support resource allocation management and decision-making resulting from the length of stay prediction.

Methods A retrospective cohort study was conducted from January 2018 to October 2020. A total cohort of 240,000 patients' medical records was collected. The data were collected exclusively for preoperative variables to accurately analyze the predictive factors impacting the postoperative length of stay. The main outcome of this study is an analysis of the length of stay (in days) after surgery until discharge. The prediction was performed with ridge regression, random forest, XGBoost, and multi-layer perceptron neural network models.

Results The XGBoost resulted in the best performance with an average error within 3 days. Moreover, we explain each feature's contribution over the XGBoost model and further display distinct predictors affecting the overall prediction outcome at the patient level. The risk factors that most importantly contributed to the stay after surgery were as follows: a direct bilirubin laboratory test, department change, calcium chloride medication, gender, and diagnosis with the removal of other organs. Our results suggest that healthcare providers take into account the risk factors such as the laboratory blood test, distributing patients, and the medication prescribed prior to the surgery.

Conclusion We successfully predicted the length of stay after surgery and provide explainable models with supporting analyses. In summary, we demonstrate the interpretation with the XGBoost model presenting insights on preoperative features and defining higher risk predictors to the length of stay outcome. Our development in explainable models supports the current in-depth knowledge for the future length of stay prediction on electronic medical records that aids the decision-making and facilitation of the operation department.

Keywords Machine learning, Healthcare management, Predictive analysis, Decision support system

†Tae Joon Jun and Young-Hak Kim contributed equally to this work.

*Correspondence:
Tae Joon Jun
taejoon@amc.seoul.kr
Full list of author information is available at the end of the article

Introduction

Background

The length of stay (LOS) conveys a crucial meaning to patients in hospitals and hospital management and clinical services. Accurate knowledge of LOS at an earlier stage will support the hospital administrators in efficiently managing bed occupancy, and improving admission and discharge scheduling alongside reducing the financial burdens to extensive cost savings in revenue [1]. All of these, allow the hospitals to make improved strategic decisions. Optimizing bed occupancy prevents patients from obtaining hospital-acquired diseases and provides physicians with greater opportunities to apply their time to patients at higher risks. Moreover, accurate LOS forecasting can aid clinicians in maximizing time from planning discharges and rerouting patients appropriately to enhance the continuity of their care [2]. Medical decision-making in providing the appropriate and best care to patients has been associated with patient preferences and growth in resource allocation [3, 4]. Overall, predicting the LOS in advance will not only benefit the hospitals both clinically and financially but enable healthcare providers to better decision-making for improved quality of care.

Related work

To reduce the LOS and improve the quality of care, prior studies have examined and attempted to accurately predict how to reduce the patient's stay. A wide variety of LOS prediction studies have been conducted in terms of population (specific patient inclusion and exclusion), outcome definition, the timeframe of a prediction, and settings in the department [5]. To illustrate, studies were conducted by populations of ICU patients [6, 7], respiratory [8], or heart failure patients [9, 10]. Other studies collected predictor variables by the time [11, 12]. From the aforementioned studies, commonly defined factors that affect LOS revealed that age, gender, comorbidities, previous historical admissions, condition at discharge, severity, and type of treatment are associated with LOS and discharge [6, 13–15].

The LOS studies were also conducted for patients undergoing an operation. Odonkor et al. proposed the need for a study after the surgery and found that age, the type of surgery, and medical histories were better LOS predictors than laboratory tests [16]. Further, Wang revealed that general anesthesia was associated with a longer LOS [17]. Alternatively, Bert used surgical patients with socio-demographic and clinical factors at pre-hospitalization visits and discovered that the age and hospitalization procedures were mainly associated with longer LOS [18]. In regards to the time frame, studies argued that the predictive ability was better captured when the model was restricted to using only preoperative predictors rather than models using preoperative factors on patients following surgery [19–23]. Substantially, most previous studies employed linear, logistic regression models in a statistical approach rather than machine learning models. Besides, how the severity of the patient's status would affect the stay after the surgery remained unknown. Therefore, to differentiate the study cohort by severity, we selected general anesthesia as an indication of severity for the reason that general anesthesia is mostly only used during major surgeries.

Here, we developed supervised predictive models implementing machine learning and deep neural network methods to adequately capture the predictive preoperative factors in a regression problem. We sought to investigate the effect of preoperative variables obtained from the patient's medical records that were associated with a LOS outcome. Our study especially focused especially on higher-risk patients by being limited to individuals who underwent general anesthesia and not being restricted to specific diseases or demographics. The novelty lies in using extensive predictive factors acquired after admission and consideration of broad cohorts, including the interpretability aspects of models to accurately investigate the predictor variables. To this end, we reviewed the electronic medical records (EMR) of 240,000 hospitalized patients admitted to the Asan Medical Center. Consequently, we have identified the twenty most important features that show a higher impact on the predictions; the six most affecting features that display the magnitude and association with the other features, and finally, four patients through which to observe the positive and negative impacts of the feature contributions. Collectively, these results will advance earlier and more precise predictions to support the operative outcomes and overall health economics.

Objectives

- To develop a predictive machine learning employed framework, which predicts the postoperative length of stay via the severe patients' preoperative data
- To support and improve the management of hospitals through 1) the decision-making process 2) the financial burden of the costs
- To interpret and evaluate the model locally and globally, thus, taking account of the individual features and patients to promote earlier risk detection alongside improving the overall treatment

Methods

Data sources

Our retrospective study was composed of electronic medical data from 240,000 patients admitted and hospitalized at Seoul Asan Medical Center in South Korea. The datasets were collected from January 1st, 2018 through October 30, 2020 (Fig. 1). Datasets for demographics, physical, diagnoses, visits, discharges, medications, laboratory, hospitalization, and operations were extracted from the ABLE data warehouse at the Asan Medical Center. The International Classification of Diseases (ICD)-10 was used to identify each patient's health condition at admission. All patients were de-identified according to the hospital's privacy rules and were randomly provided a unique patient ID, which acted as a key linkage between the datasets.

Cohort preparation and outcome definition

The patient population was initially defined by filtering the population with general anesthesia during the operation in index Datetime, defined as 'indexOPDT'. This criterion was chosen because general anesthesia is often used for a wide range of surgical procedures, including those that are more severe or complex. This makes it a relevant indicator for capturing a diverse patient population, allowing for a comprehensive analysis of factors affecting Length of Stay. The measure of the patient's severity pre-operation was indicated based on whether the patient required general anesthesia. The inclusion criteria for the patient population were as follows: 1) the existence of 'indexOPDT' and 'LOS outcome' values, 2) alive before the operation date 3) provided general anesthesia, 4) the duration of the length of stay period was greater than three but less than thirty days. These criteria were set to focus on patients with a typical range of hospital stays, excluding extreme outliers that could skew the analysis. Additionally, exclusion criteria included patients whose death date was prior to the operation date and patients who undertook anesthesia other than general. These excluded patients did not provide the LOS outcome variable. The flowchart in Fig. 2 displays formerly described cohort preparation. The prediction outcome of our retrospective study was continuous, in days. The length of stay was defined as the period from the date of the first operation until the date of discharge.

Data collection

The electronic medical record of each patient's history was distinguished by variables collected from demographics (age, gender, cancer, death dates, and admission date), physical (BMI, BSA, weight, and height), diagnosis (ICD-10 diagnosis code), visit (ward type, department, path, and visit date), discharge (discharge date), medication (medication name), laboratory (laboratory test date, laboratory result, and laboratory test name), hospitalization (department transfer, and diagnosed department), operation datasets (operation datetime, age at operation, operation department, operation procedure, operation type, ICD-9 surgery code, and anesthesia code). Patient diagnoses used ICD Tenth Revision (ICD-10) codes, obtained from the patients at the hospital admission. A total of 422 variables (excluding the patient id and the outcome variables) were included as preoperative predictors for each patient. The predictive factors were mainly selected with the support of cardiologists, who offered assistance in deciding the importance and suitability of the variable.

Model

This section discusses the overall process of model construction. The prediction model was built in eight steps (Fig. 3).

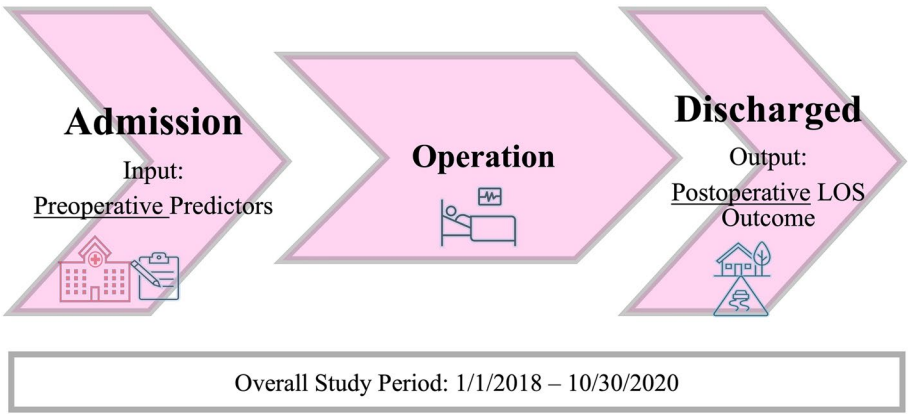


Fig. 1 Illustration of the study period

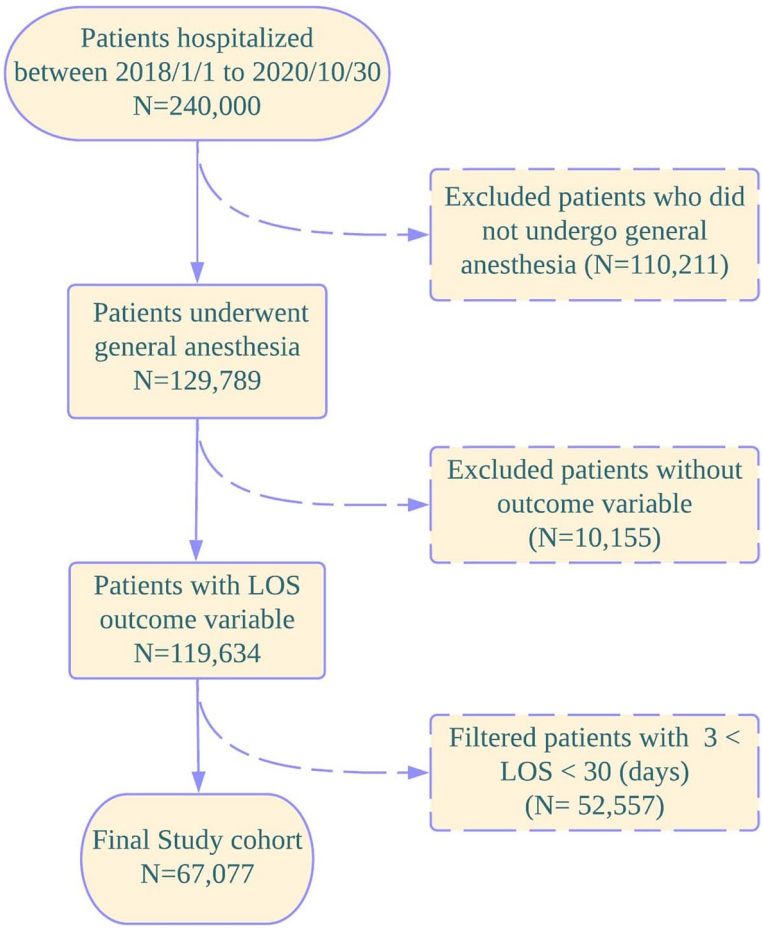


Fig. 2 Flow chart of the patients included in the study. The patient exclusion was conducted if the patient died before the operation, or a patient undertook anesthesia other than general, thereby, not providing the LOS outcome variable

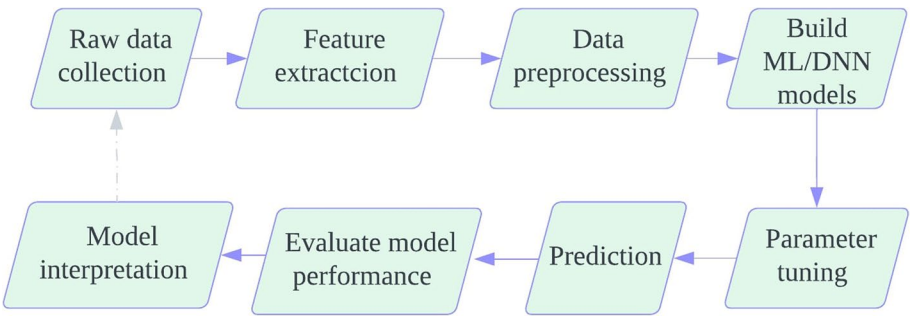


Fig. 3 Flowchart of detailing how the prediction model in the postoperative length of stay was built

Data preprocessing and feature engineering

To create increased performance results, the data-preprocessing stage is essential. During the preparation step, raw data are examined with categorical data, outliers, missing values, or redundancy [24, 25].

To handle the categorical values, we used one-hot encoding for nominal variables to avoid implying any ordinal relationship, and label encoding for ordinal variables to preserve their order. Binary encoding was initially used for some variables, such as gender (male = 0,

female = 1), cancer and death dates (occurrence = 1, none = 0), operation occurrence (exists = 1, none = 0) and department transfer (yes = 1, no = 0). However, to address the concern that the model might interpret these binary encoded variables as numeric, we have revised our approach. Instead of binary encoding, we have used one-hot encoding for all categorical variables, including binary categories, to ensure the model does not assume any numeric relationship. A large number of observations in categorical features such as diagnosis and medication names were filtered with a threshold of the hundred most frequent observations. Outliers for height (cm) and weight (kg) variables were removed. Heights below 100 cm and above 250 cm were excluded from the analysis, as well as weights below 30 kg and above 200 kg, to ensure data accuracy.

To enhance the performance of the model prediction, missing data were handled individually for each dataset. Firstly, the features missing more than eighty percent were eliminated. This threshold was chosen based on a balance between retaining as much data as possible while excluding features that would provide minimal information due to their high rate of missing values. Additionally, we performed multiple imputation methods, including mean and median replacement, and conducted a sensitivity analysis to ensure the robustness of our results. Secondly, the null values for BMI and weight were filled with the median to account for their skewed distributions, whereas the mean was used for height, which was more normally distributed. Further, the overall missing data were imputed using the k-nearest neighbors (KNN) method to preserve the data structure without implying any physical meaning. This approach was chosen over imputing 0 or -1 to avoid any misinterpretation by the model. Lastly, before the dataset is learned with models, we used standardization by standard deviation to scale the data, ensuring that the explainability of SHAP is not affected by non-linear transformations.

Feature selection

Feature selection is a substantial process of identifying a subset of features from the input data that consequently reduces the irrelevant features, and training time, and reduces dimensionality, thereby, supporting better prediction performances in machine learning [26]. Initially, the study started with 500 features. After applying the ANOVA F-value feature selection, the feature set was reduced to 300 features. While modern machine learning methods such as XGBoost, RF, Ridge regression and deep neural networks can handle high dimensional data, feature selection was employed to improve model interpretability, reduce overfitting, and enhance computational efficiency. By selecting the most relevant subset of features, we aimed to enhance model's ability to generalize the unseen data. We identified and selected the most relevant subset of features by ranking them based on their ANOVA F-values. The selection process involved iteratively testing different values of k (the number of top features) to find the optimal number of features that maximize model performance. The algorithm selected features based on the highest ANOVA F-values. The value of k was chosen from a range of 10 to 200, with the optimal k of 20 ultimately being selected for the model. This method, while computationally efficient, significantly impacts defining the predictors and ensures that only the most informative features are used in the model. Furthermore, the model was augmented by incorporating additional features derived from the most significant predictors identified through feature importance analysis. This included creating interaction terms and polynomial features to capture complex relationships within the data. The feature selection and augmentation processes collectively aimed to balance model complexity, interpretability, and predictive performance.

Model construction and experimental setup

We followed a regression approach for the continuous LOS outcome variable prediction and employed four types of learning algorithms to develop the predictive models: (1) ridge regression [27], (2) extreme gradient boosting (XGBoost) [28], (3) random forest [29], and (4) multilayer perceptron (MLP) [30].

The ridge regression is an extended technique of the linear regression algorithm that uses L2 regularization with a squared magnitude of coefficient and lambda factors added as a penalty to the loss function. Consequently, the ridge estimator reduces the variance and shrinks the coefficients toward zero.

The XGBoost is a highly scalable, gradient tree-based ensemble system that parallelly runs at a much faster rate compared to existing supervised machine learning techniques. XGBoost is effective as it minimizes regularizations (L1 and L2) to loss function and handles missing data using a sparse approach. Additionally, the iterative training process of inserting newly created trees combined with previously built trees ultimately minimizes the loss.

Another tree-based mode, RF is an ensemble algorithm that uses the extended technique of bagging, and feature randomness, to create a subset of separated observations. Hence, it is great for high dimensional data and quickly trains the features to maximize the information gain [31, 32].

The MLP is a fully connected feedforward artificial neural network for supervised learning and as the name refers, one or more hidden layers of activation functions

exist in the middle. It uses a backpropagation technique [33] that has a nonlinear function to optimize the weights and reduce the prediction errors. A study found that MLP provided the best performance result among other machine learning models [34].

To perform data sampling, the dataset was randomly split into a training and test set by an 80:20 ratio. The training sets were used to learn the mapping of inputs whereas the test sets were used to evaluate the predictive ability of the model. Following the data stratification, the input features on the training dataset were normalized in the range of zero to one. All four models were built with the same predictor variables through which the model development was performed using the open-source Python package.

Parameter and hyperparameter tuning

To control and improve the model's performances, we jointly used randomized and grid searches to achieve the optimal combination of the hyperparameters for the models. The search iterates over the previously defined sets of hyperparameters to obtain the best-tuned values. Then, a 3-folder cross-validation is used to evaluate each set of hyperparameters into the folds, which the model subsequently fits on to ultimately provide the score on the combination values.

To inform the detailed parameter tuning, for the ridge model, a penalty of 1 is used as a constant that controls the regularization strength.

The random forest model used 100 trees to fit the model that intakes one feature at a time, and a minimum of two internal node samples that will hold before splitting to a further node. Also, a minimum of 1 sample is required to be a node.

To train a multi-layer perceptron layer model, we first provided a specification for an input vector containing 530 features. Following the input layer, a further three hidden layers and the output layer are created with ReLU activation. Moreover, each layer (hidden and output) is created with output neurons of 20, 10, 5, and 1, respectively. Amid the layers, a 25 percent dropout layer is added after the second hidden layer. The MLP model is compiled with the adam optimizer at 0.0001. For evaluation metrics, the mean-squared error loss argument and the root-mean-squared error are defined. The established model and compiled information are trained with 30 epochs, a batch size of 200, and 15% of the training data, which is used for validation. Consequently, six layers with a sum of 8,731 parameters were trained in hidden units during the fitting.

The XGBoost model was created with 150 trees used in the forest with five maximum depths of tree learned at 0.1 rates. Also, 1 feature is randomly subsampled to train a new tree.

All four models are created with the best setting, which was chosen according to the model's performance. The parameters are tuned with the Grid Search CV package from an open-source Sklearn model selection, while the MLP model implementation was built with the Tensorflow, Keras library.

Evaluation

In this research, the RMSE (root-mean-squared error) metric was used to evaluate the model performance in measuring each of the regression model's errors. The errors signify, on average, the margin of error between the predictions and the true target values. Therefore, the units are represented by the original units of the predicted value, which in our study are 'days'. The RMSE was measured for both the training and test sets to verify the closeness of the evaluation results. After evaluating the RMSE results from four different models, we further analyzed and interpreted the XGBoost model which demonstrated the highest prediction performance. For the MLP model, RMSE was estimated through the loss function for the LOS outcome in days.

Model interpretation

SHAP (SHapley Additive exPlanations) is a Shapley-based novel approach based on game theory, which aids a better understanding of the tree-based model structure for each prediction [35, 36]. The explanation measures the interactions between the local features which then combines to deliver insights into individual predictors and the whole model. SHAP also enables the characterization of the high and low-risk predictors and identifies predictors that degrade over the model's performances. We employed SHAP values to improve the interpretation of the learning results of the XGBoost model, and to further analyze the associations between the features that identify the critical predictors and impact the LOS outcome. The SHAP tools we used includes dependence plots, summary plots, and force plots.

Results

Baseline characteristics of participants

Among the 240,000 hospitalized patients admitted to Asan Medical Center, 67,077 patients had undergone general anesthetic procedures were included in this study. The patients were randomly stratified to the training dataset ($n=53,661$) and test dataset ($n=13,416$). The mean age was 57.0 years and included 29,608 males alongside 37,469 females. The mean LOS outcome for the severe patients was 6.7 days (s.d. = 4.7 days). A summary of baseline patient characteristics is presented in Table 1.

Table 1 Baseline patient characteristics

Variable	Number of encounters	Training	Test
Cohort characteristics	Total *n* = 67,077	Total *n* = 53,661	Total *n* = 13,416
Demographic			
Age (years), mean (s.d)	57.0 ± 14.8	57.0 ± 14.7	56.9 ± 14.9
Gender, n(%)			
Male	29,608 (44.1)	23,386 (43.6)	5,934 (44.2)
Female	37,469 (55.9)	30,275 (56.4)	7,482 (55.8)
Body mass index (kg/m2), mean (s.d)	24.5 ± 3.8	24.5 ± 3.8	24.5 ± 3.8
Disease characteristics			
Cancer, n(%)	866 (1.3)	688 (1.3)	178 (1.3)
Hypertension, n(%)	13,962 (20.8)	13,015 (24.3)	3,750 (28.0)
Diabetic mellitus, n(%)	6,068 (9.0)	6,068 (9.1)	1,194 (8.9)
Liver disease, n(%)	2,282 (3.4)	1,761 (3.3)	383 (2.9)
Renal disease, n(%)	843 (1.3)	690 (1.3)	162 (1.2)
In-hospital death, n(%)	515 (0.8)	413(0.8)	102 (0.8)
Visit path			
Outpatient- immediate	2,734 (4.1)	2,186 (4.1)	548 (4.1)
Outpatient- reservation	55,467 (82.7)	44,425 (82.8)	11,042 (82.3)
Inpatient	12,956 (19.3)	10,374 (19.3)	2,585 (19.2)
Emergency Room	5,868 (8.7)	4,676 (8.7)	1,192 (8.9)
Department transfer, n(%)	5,967 (8.9)	4,764 (8.9)	1,203 (9.0)
Length of stay after surgery, mean (s.d.)	6.7 ± 4.7	6.7 ± 4.7	6.7 ± 4.6

We further examined the distribution of the outcome variable, the length of stay by the total number of patients and by the department. Figure 4 demonstrates the overall distribution of the patients from LOS days 3 to 30. Expectantly, the trend depicts the length of stay at the hospital decreasing sequentially as the number of days increases. Moreover, Fig. 5 illustrates the distribution of LOS over the departments that patients first visited during the hospital visit. The result indicates that the liver transplant surgery department had the most LOS days. In

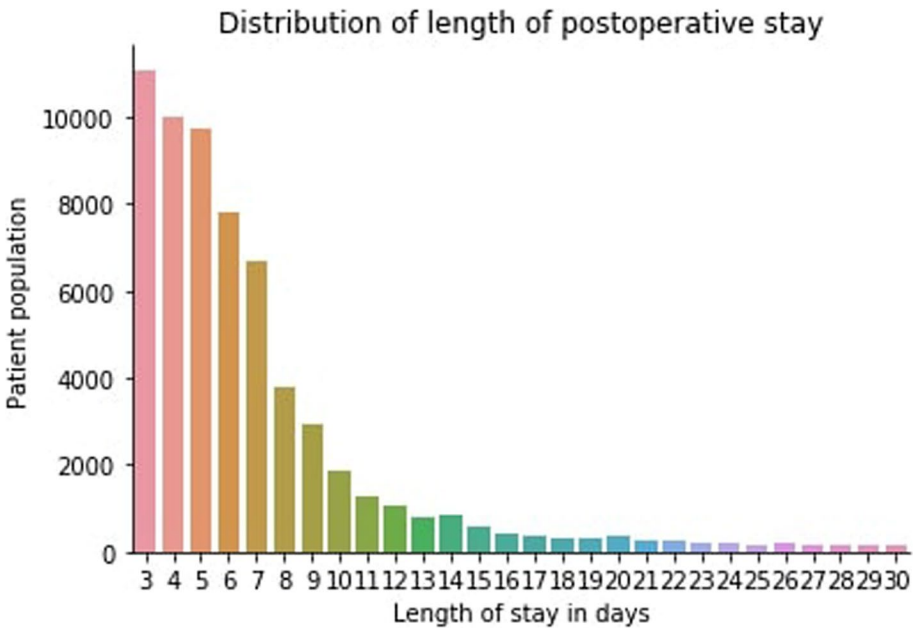


Fig. 4 Distribution of LOS by the total number of patients. The y-axis presents the number of patients versus the LOS in days displayed on the x-axis

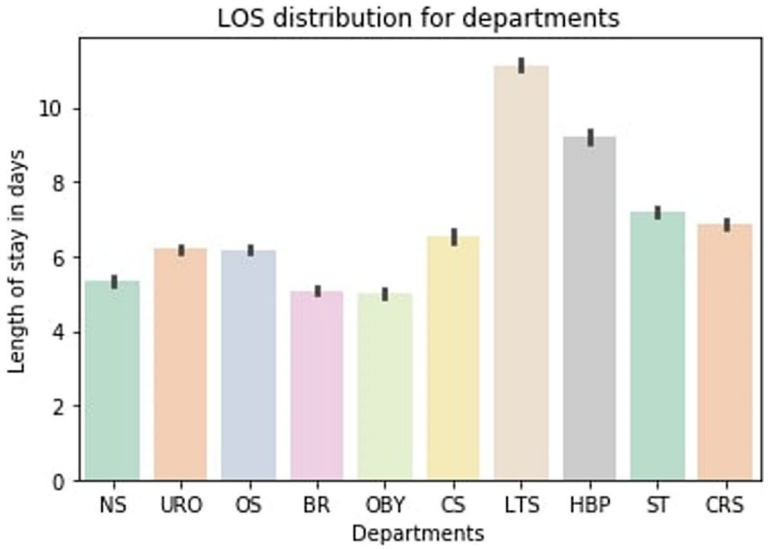


Fig. 5 Distribution of LOS by the department. The order of department from left to right corresponds to the followings department, respectively: Neurosurgery, urology, orthopedics, breast surgery, obstetrics and gynecology, cardiovascular and thoracic surgery, liver transplant surgery, hepatobiliary and pancreatic surgery, gastrointestinal surgery, colorectal and anal surgery

contrast, patients spent the least amount of LOS days in the breast surgery and obstetrics and gynecology departments (Fig. 5).

Model evaluation

In total, 422 preoperative features were extracted. Using the aforementioned and described models (Methods section), we explored four types of regression prediction models with the extracted features from individual patients and evaluated their performance through the RMSE, MAE, MAPE, and R^2. These additional metrics provide a more comprehensive evaluation of the model's predictive performance. We also divided the test set into five buckets based on the length of stay (0–20%, 20–40%, 40%-60%, 80–100%) and reported RMSE for each bucket. This provides a more detailed evaluation of model performance across different patient groups. The Ridge linear regression model obtained a 3.72, XGBoost provided a 3.56, MLP of 3.71, and RF produced an RMSE of 3.64 (Table 2). The XGBoost model generated the best performance with the lowest RMSE evaluation score. Additionally, XGBoost demonstrated superior MAE, MAPE, and R^2 values, indicating its robustness and accuracy in predicting the length of stay.

We further conducted hyperparameter tuning using grid and random searches to optimize the model's performance. Table 3 presents the minimum and maximum values tested for each hyperparameter, along with the range of variation, used in the grid and random search for hyperparameter tuning of the Ridge, RF, MLP, XGB models.

Analysis of the individual predictive feature

We exploited the SHAP method on the XGBoost model to further our interpretation of its principal performance. The SHAP model automatically chooses a few of the 422 features from the training data input that gives the impact and risk outcomes at the specific observation time [Lundberg]. The summary plot in Fig. 6 represents the directionality of individual features' contribution to the model prediction. The x-axis presents the mean absolute SHAP value and indicates the positive and negative

Table 2 Performance comparisons of the model predictions. The root means square error metric is used for evaluation

Model	RMSE (0–20%)	RMSE (20–40%)	RMSE (40–60%)	RMSE (60–80%)	RMSE (80–100%)	RMSE	MAE	MAPE	R squared
Ridge	2.10	2.50	3.00	3.80	4.10	3.72	2.90	13.5	0.85
XGBoost	2.00	2.40	2.80	3.60	4.0	3.56	2.72	12.8	0.87
MLP	2.15	2.55	2.95	3.75	4.05	3.71	2.84	13.2	0.85
RF	2.05	2.85	2.85	3.65	4.05	3.64	2.80	13.0	0.86

Table 3 Hyperparameter Ranges Tested for Each Model. Minimum and maximum values tested for each hyperparameter tuning of the Ridge, RF, MLP, and XGB models

Model	Hyperparameter	Minimum Value	Maximum Value	Range
Ridge	Alpha	0.01	10	0.01
RF	Number of Trees	10	500	10
	Max Depth	3	50	1
MLP	Hidden Layers	1	10	1
	Neurons per Layer	10	300	10
XGB	Learning Rate	0.001	0.3	0.001
	Number of Trees	10	500	10
	Max Depth	3	50	1

impact of the features on the target. The y-axis lists the top twenty most impacted features selected and ranked in descending order to indicate the feature's importance. Each point of a row from the summary plot in Fig. 6 (a) represents a feature from the test set. The color indicates whether the variable is high or low for the observation. The global SHAP analysis revealed that several preoperative medication data points were among the top predictors of LOS. These insights suggest that clinicians should carefully monitor and manage medications prior to surgery, as this can significantly influence recovery times. For example, the SHAP values associated with bilirubin levels and albumin indicate that optimizing these biomarkers before surgery could potentially reduce extended stays. This kind of actionable insight provides a practical way for clinicians to improve patient outcomes. High levels of direct bilirubin, a departmental change, and the removal of other organ, citric acid, albumin, and mupirocin content contributed to a high and positive impact on the quality rating. Alternatively, an increased presence at the department of cardiovascular and thoracic surgery during their visit promoted a high and negative impact on the quality rating. Figure 6 (b) shows the overall contribution of the top 20 important features.

To further analyze the importance of features from the SHAP feature interpretation, we examined each of the six most importantly affecting features with dependence plots (Fig. 7). All features other than the direct bilirubin laboratory test and departmental change feature were binary encoded, thus, illustrating the SHAP values between zero and one. The binary features show a trend whereby there are more ones than zero values. Conversely, direct bilirubin, a continuous feature, possesses SHAP values within the range of 0 to 0.7, which shows the minimum and maximum values, while the values predominantly remain in the range of less than 0.1. The department change feature depicts three labels encoded as 0, 0.5, and 1, showing the most at 0.5. The dependence plot also presents the

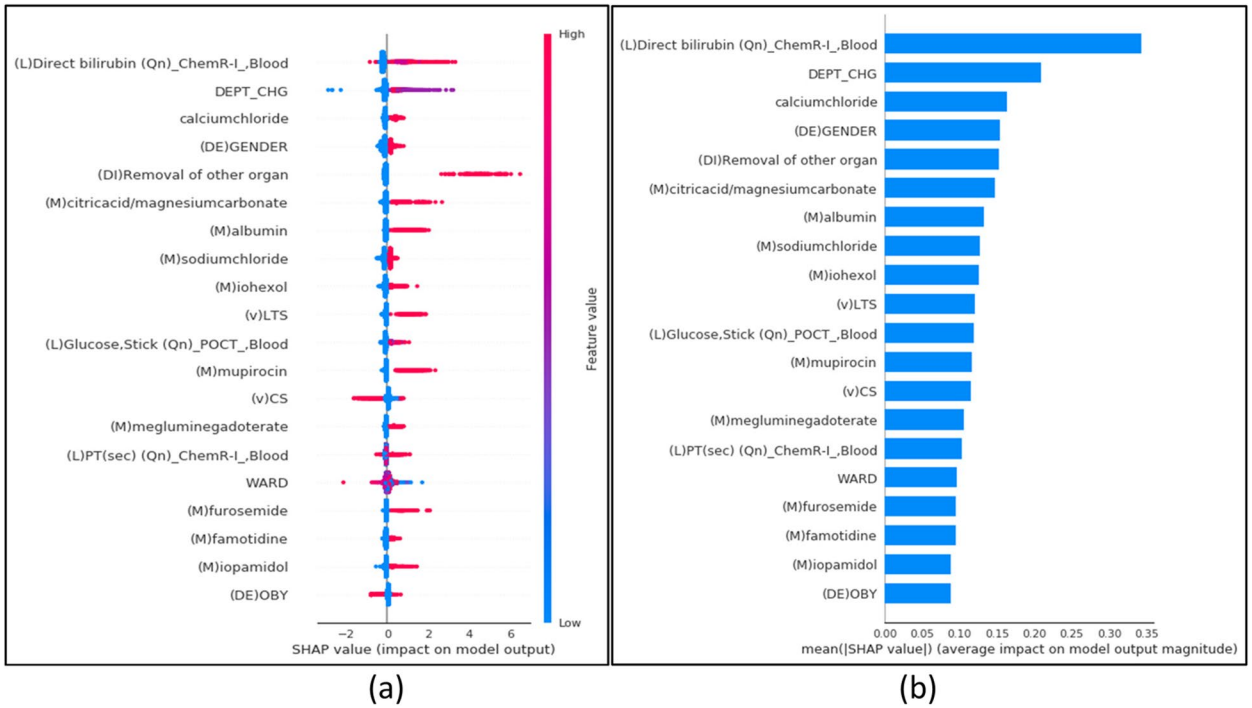


Fig. 6 The SHAP summary plots for model interpretation. The top 20 important features are represented in the order of the highest to the lowest mean absolute value of SHAP values. **a** The summary plot shows the sum of SHAP values and each predictor's impact on the overall model prediction performance (**b**) The bar plot results are based on the mean value of the SHAP values

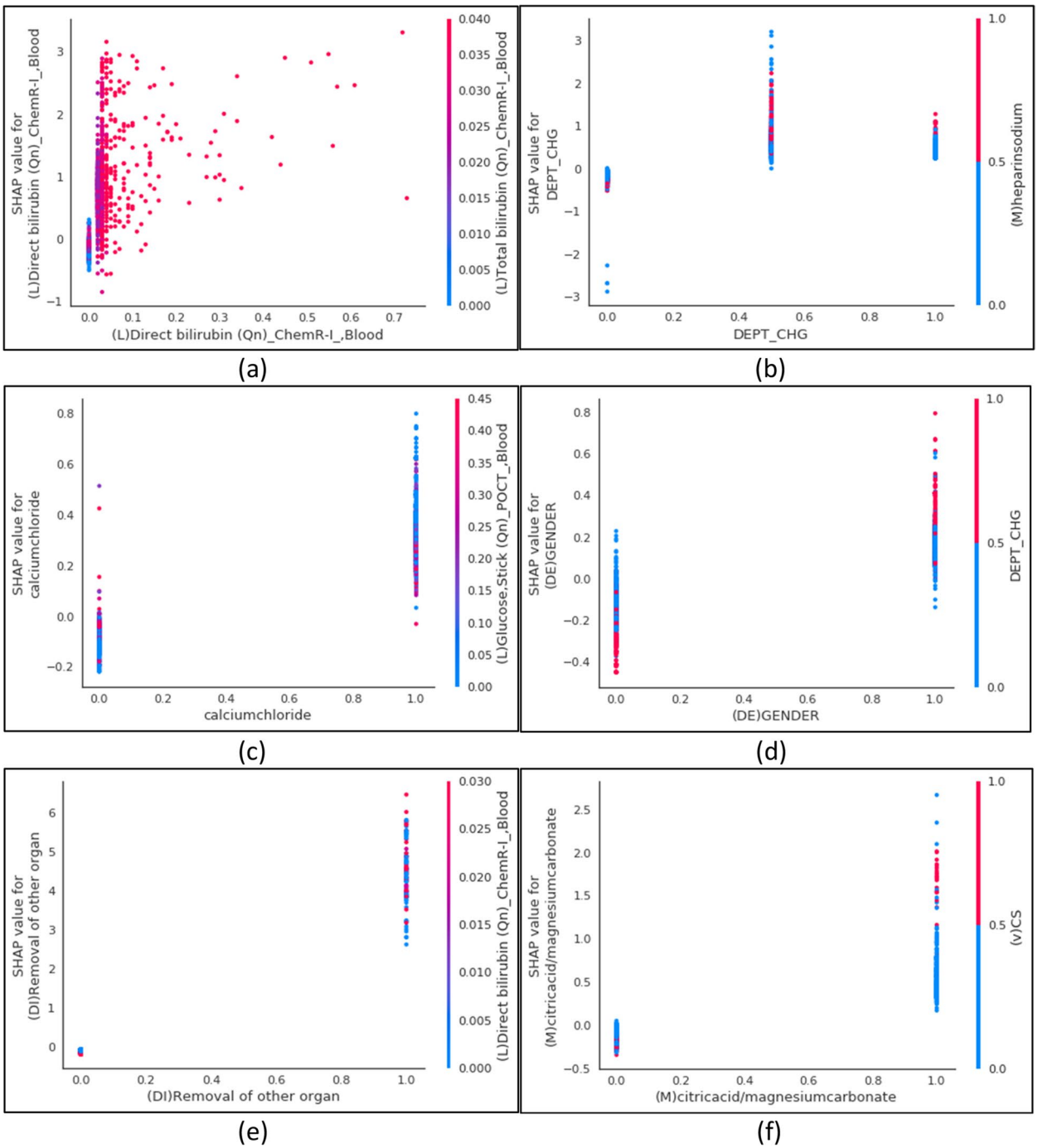


Fig. 7 The SHAP dependence plots. The plot displays the variable that interacted with the target. The x-axis represents the feature value and the y-axis represents the SHAP value of the feature. The color corresponds to the value of the interacted feature. **a** The direct bilirubin blood laboratory test predominantly interacts with the total bilirubin blood laboratory test. **b** The department change mostly interacts with heparin sodium medication. **c** The calcium chloride medication primarily interacts with blood glucose stick laboratory tests. **d** Gender mainly interacts with department change. **e** The removal of other organ diagnoses mostly interacts with the direct bilirubin blood laboratory test. (**f**) The citric acid and magnesium carbonate medication principally interacts with the cardiothoracic surgery department

feature with the strongest interaction. For instance, gender is reported to mostly interact with departmental change where the females are coded as 1 and males as 0 (Fig. 7-d). The probability of the LOS risk is more likely to be predicted in males who have undertaken a departmental change than the males who have not. Interestingly, these parameters are inverse for females who change departments, with the probability of the LOS being more likely to be predicted following a department change than without. Moreover, the dispersion of SHAP values may show the iteration between the two continuous features (Fig. 7-a). The direct bilirubin is captured to interact with the total bilirubin. Indeed, there is higher interaction between the two features when the direct bilirubin values are less than 0.1. Moreover, as the direct bilirubin values increase the observed interaction with the total bilirubin diminishes.

Moreover, to deepen our understanding of the features of the model performance, we created four individual observational force plots. The force plot depicts explanations for each feature contributing to the model output from the base mean value of train data. Each plot shows a set of patient features from a single input feature of the dataset. The base value and the expected value presents the model output average from the training dataset. The output value presents the prediction for observation. The order of features listed in each plot indicates the size of the contribution. The risk parameter explains the features in red elevate the risk factor of the prediction, whereas the features in blue reduce it to a lower risk.

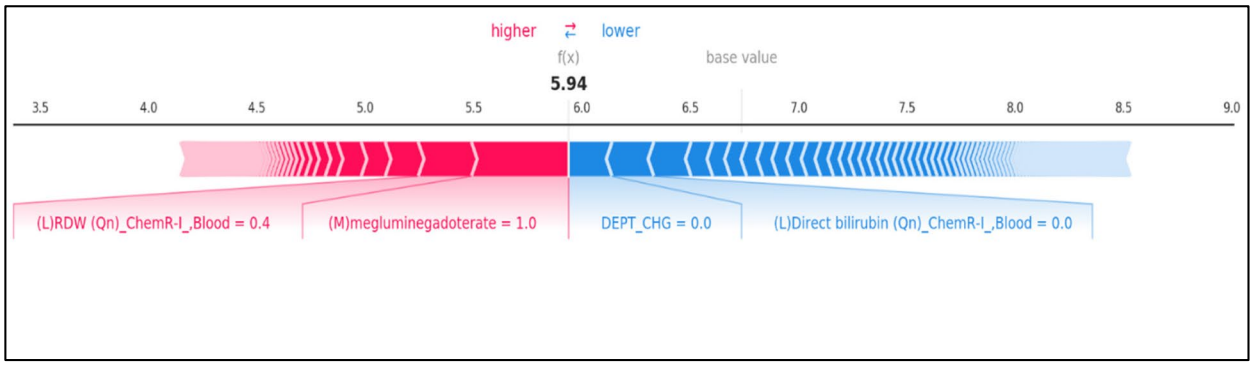
The observation from the first patient's records in Fig. 8-a illustrates that both the meglumine gadoterate medication and RDW laboratory test provide a positive contribution. However, the departmental change and direct bilirubin laboratory test factors have a negative contribution. The meglumine gadoterate and department change are the primary positive and negative contributing variables, respectively. The total negative contribution is smaller than the positive contribution, therefore, producing an overall smaller output value than the base value. Contrastingly, the observations made for a patient in row 201(Fig. 8-b), belonged to the true positive group implying a high probability of LOS from the patient's prediction. The result shows that the direct bilirubin test, malignant neoplasm of extrahepatic bile duct diagnosis, and ranitidine medicine treatment promote a higher prediction of LOS prediction. Overall, the patient from row 501 (Fig. 8-c) provides the highest prediction, as the output value and the base value are the closest to each other here.

Discussion

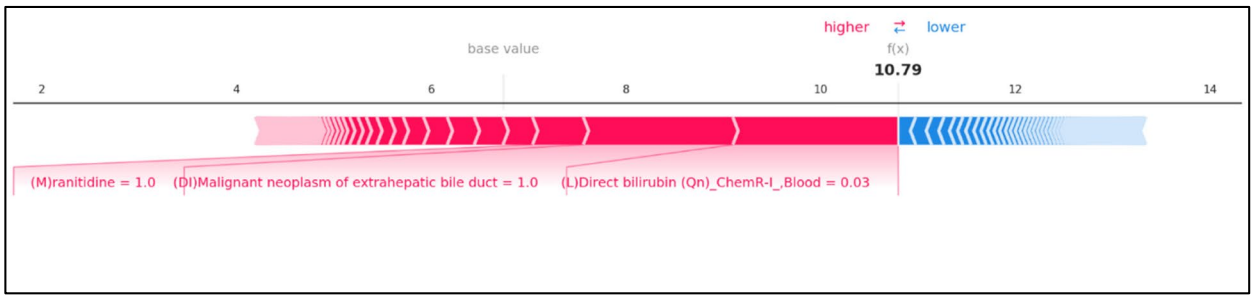
This research focused on developing machine learning and deep neural network-based models to predict the length of stay (LOS) for severely at-risk patients undergoing surgery with general anesthesia. Among ridge regression, XGBoost, multi-layer perceptron neural networks, and random forest models, XGBoost emerged as the best-performing model, demonstrating its utility in this clinical context where traditional models might struggle to capture complex interactions. Accurate LOS predictions empower hospital administrators to optimize resource allocation, manage bed availability, and schedule critical care units, particularly during peak periods.

Moreover, our study provides critical insights into the application of machine learning models, particularly XGBoost, to predict LOS for severely at-risk patients undergoing surgeries with general anesthesia. By analyzing 422 preoperative features across 67,077 observations, XGBoost consistently outperformed other models, making it a robust tool for forecasting postoperative LOS. On average, our model predicted that severe patients would stay 3.56 days in the hospital post-operation. These predictions allow hospital administrators to proactively plan and allocate resources efficiently based on patient needs. Specifically, our model identified that most LOS for these patients clustered between 3 to 7 days, allowing hospital management to anticipate and optimize resources for peak periods. When the model predicts a longer LOS, clinical teams can adjust treatment protocols or initiate post-op rehabilitation earlier, potentially shortening recovery times and improving patient outcomes by addressing risks preoperatively.

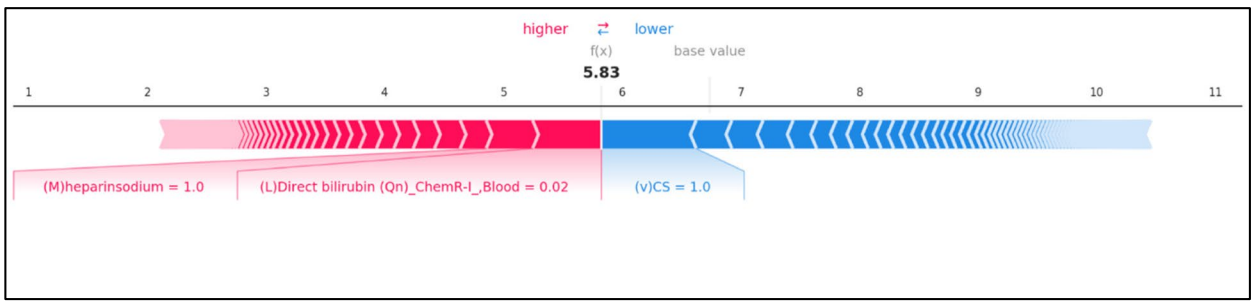
Furthermore, the analysis suggests that hospitals can proactively allocate resources more efficiently by predicting LOS with XGBoost. The population distribution plot indicated that most patients have a stay duration concentrated between 3 to 7 days. With this knowledge, hospital management could plan for peak times, ensuring that adequate resources and staff are available to manage crowding. Additionally, the longest LOS predictions, such as those for liver transplant surgeries, highlight departments that may need extra resources for post-operative care. These predictions will ultimately lead to shortening the overall duration outcomes. In healthcare, decision support systems are significantly important to both the providers and the patients [37, 38]. Effectively reducing time management by supporting the decision-making for clinicians will eventually provide high-quality care and drive the right clinical outcomes for the patients [39, 40]. By providing accurate predictions for LOS, the model helps avoid unnecessary extended stays and reduces the risk of hospital-acquired complications and unplanned readmissions. This leads to significant cost savings as fewer resources are wasted on prolonged hospital stays, and patients are discharged on time. These efficiencies free up beds and resources for more patients, improving hospital throughput and



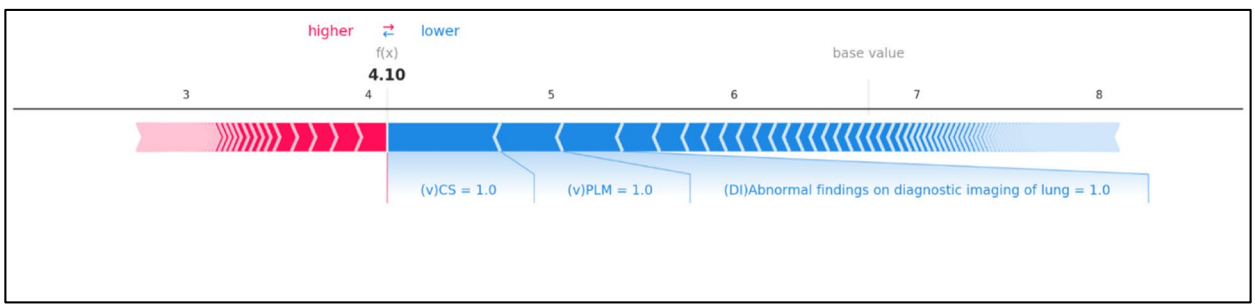
(a) A patient in the first row provides a prediction of 5.94.



(b) A patient in row 201 produces a prediction of 10.79.



(c) A patient in row 501 yields a prediction of 5.83.



(d) A patient in row 5001 supplies a prediction of 4.10.

Fig. 8 SHAP force plots. **a** A patient in the first row provides a prediction of 5.94. **b** A patient in row 201 produces a prediction of 10.79. **c** A patient in row 501 yields a prediction of 5.83. **d** A patient in row 5001 supplies a prediction of 4.10

enhancing revenue optimization. Moreover, by aligning staffing needs with real-time LOS predictions, hospitals can prevent overstaffing and reduce labor costs. Integrating LOS predictions into hospital decision-support systems allows for real-time adjustments to surgical schedules, staffing, and resource distribution, improving

patient care while reducing operational inefficiencies and financial burdens. Shameer et al. (2017) demonstrated that predictive models could significantly cut costs by guiding the allocation of intervention resources for surgical readmissions. Our findings align with such studies, showing that predictive models not only forecast LOS but also play a pivotal role in hospital resource management and cost reduction [41].

For instance, Rath et al. (2017) developed a quantile prediction model to predict surgery duration, which is then used within an optimization model for scheduling surgeries [42]. Additionally, Misic et al. (2021) quantified the cost savings from using simulated models to guide intervention resources for hospital readmissions [43]. Including such considerations in the development and application of LOS prediction models can provide significant benefits to healthcare operations and patient care outcomes. Therefore, implementing an accurate but deeper analysis system of interpreting the risk predictors to improve the decision-making prior to an event plays a crucial role in anticipating the patient's stay. In this respect, in agreement with Bertsimas's findings [44], we demonstrate that by adding explainability to machine learning results the analysis in predicting risk factors for delays in discharge can be further enhanced. To accomplish this, SHAP analyzes the model learning more accurately and consistently through global and local explanation approaches to further investigate our model.

Initially, the global explanation approach presented the top twenty preoperative predictors contributing to forecasting the LOS outcome for critical individuals (Fig. 6). These provide sound evidence of the SHAP technique being able to sufficiently provide clinical importance through the use of the XGBoost model. Among the top twenty most influential features, ten emerged from the medication data points relating to the severe patients' prescription prior to the surgery being predominantly associated with the subsequent LOS. Moreover, this could lead to the physicians taking a high level of selected features as risk factors prior to arranging the pre-surgery procedures. We found similar key markers for predicting the LOS as previously detailed by Iwase et al. and co-authors', where albumin and direct bilirubin were associated with the LOS for critically ill patients. Especially, the observed magnitude of direct bilirubin test attribution, with lower than 0.1 units (Fig. 7-a), alongside the removal of other organs diagnoses indicates an association. Furthermore, these defined predictors will reduce the discrepancy in selecting the irrelevant features pre-surgery providing only the essential resources to the physicians. Thus, to effectively reduce the hospitalization days, we suggest four-fold implications: Firstly, a patient's contribution to avoiding the pre-surgery medication intake according to the lists presented by our findings; secondly, physicians attentively devote more attention to specific blood laboratory results paired with the patient's prescribed medication (Fig. 7); thirdly, the hospital operational team strategically manages patients receiving treatment at certain departments and avoids additional, impractical departmental changes, particularly for females (Fig. 7-d).

Furthermore, the knowledge gained from the local explanation approach (Fig. 8), affirms that patients may also progress into hospital management. Thus, results in financial improvements since the discovered features for each patient can detail the contribution from the whole model output. Primarily, the operations management team in a hospital may take the findings of these positively contributing preoperative predictors from patients, which relate to true positive prediction performance and accurately predict the stay (Fig. 8-c). To a greater extent, these patient-inspired findings present an opportunity to perform patient-specific care, whereby individuals are encouraged to engage with their medication to reduce unnecessary healthcare costs and inefficient clinical trials, to ultimately shorten their postoperative stays [45]. Nevertheless, because the most impactful predictors display variances between the patient observations, the experiment should be conducted globally to further refine the cohorts.

Using the SHAP interpretation, our study achieved better decision-making from patients' visits by detecting the risk factors coupled with the predictors. Overall, based on our feature identification, the LOS could be more highly and accurately designated at an earlier stage of the treatment process (Figs. 6, 7). In summary, an accurate analysis of the importance and contribution of the XGBoost model's preoperative predictors to the operative LOS will both support the facilitation of the operation department and provide efficient resource allocation toward advancing overall hospital management.

Previous studies have developed algorithms for predicting the LOS focused on disease-specific surgeries [46–49]. Predicting the risk factors of critically ill patients is significant since a prolonged stay for patients can increase the risk of hospital-acquired infections and hinder other patients' access to the operation and medical resources [50]. A longer LOS is reported to be related to the illness severity [51]. Additionally, Naessens proposed higher-risk populations are likely to incorporate considerably more resources [52]. Hence, we focused on cohorts with critical patients, narrowing the focus to patients who had undergone general anesthesia. Consequently, our study offers substantial contributions to hospital management, operational efficiency, and clinician decision-making. The superior performance of

our XGBoost model, enhanced by SHAP explainability, allows for more precise and actionable predictions of postoperative length of stay (LOS). These predictions enable hospital administrators to optimize resource allocation, manage bed availability, and streamline patient flow, reducing bottlenecks and improving overall operational efficiency. For instance, accurate forecasting of LOS can inform staffing needs and surgical scheduling, ensuring that the right resources are in place at the right time, especially during peak periods or in high-demand units such as critical care or liver transplant surgery departments.

Furthermore, clinicians can use this model to assess the risk profile of each patient and plan individualized treatment approaches. For instance, if the model predicts a longer LOS based on preoperative factors like elevated bilirubin levels or the patient's medication history, clinicians can intervene early by optimizing treatment protocols, adjusting medications, or scheduling closer monitoring post-surgery. This approach not only improves patient outcomes but also reduces the likelihood of complications during recovery. This empowers healthcare providers to make more informed, data-driven decisions that can lead to targeted preoperative interventions, ultimately improving patient outcomes. The ability to tailor care based on individual risk factors enhances not only clinical decision-making but also patient engagement in their own recovery process, leading to more efficient postoperative care and reduced hospital stay durations. These advancements, grounded in both predictive accuracy and model explainability, can lead to significant cost reductions, improved resource utilization, and more effective patient care strategies across healthcare systems.

Limitations

Yet, our study contains several limitations. Firstly, the predictive model developed in this study is specific to the facility where it was developed. The model's applicability to other facilities, especially those in different countries with varying disease compositions and medical systems, is limited. Therefore, the developed model can only be used for medical management within the model-developing institution. This single-centered data analysis may limit any further validation from external resources. Future works should consider taking the external validation development from multiple sites to enhance the model's predictive performance. Secondly, no socio-economic and behavioral data were included in the study, which could have impacted postoperative LOS. Future studies should include a gender-based analysis to identify any potential disparities in prediction accuracy and ensure fairness across demographic groups. Additionally, incorporating socio-economic and genetic data could further enhance the model's precision by accounting for individual patient circumstances, leading to more personalized healthcare strategies [53, 54].

Conclusion

To our knowledge, this retrospective study is the first to consider all the following aspects: 1) to explore postoperative LOS predictions for non-disease specific but particular to the severity of the patient's condition: surgeries that require general anesthesia, 2) to evaluate further analysis with SHAP interpretation both locally and globally on the model along with the individual features, 3) to fully incorporate the EMR of large-sized cohorts and features, highlighting the potential importance of features analyzed in this study. The model's forecasting ability and predictor factors indicated in this study will further support the utilization of the resource allocation systems in hospital management, decision-support for health providers, and patients to promote engagement in their healthcare journeys.

Abbreviations

LOS	Length of stay
EMR	Electronic medical records
ICD	International classification of diseases
XGBoost	Extreme gradient boosting
MLP	Multilayer perceptron
RMSE	Root-mean-squared error
SHAP	SHapley Additive exPlanations

Acknowledgements

This work was supported by the Korea Medical Device Development Fund grant funded by the Korean government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, Republic of Korea, the Ministry of Food and Drug Safety) [Project Number: 202012B06]; and by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health Welfare, Republic of Korea [grant number: HR21C0198]. The funders of the study had no role in the study's design, data collection, data analyses, data interpretation, or writing the manuscript. All authors had full access to the data in the study and accept the responsibility to submit it for publication.

Authors' contributions

H.C.: research design, analysis, model development, and writing; I.A., H.G., H.K., Y.K., H.S., H.C., M.K., J.H., G.K.,S.P.: research design, review; T.J., Y.K.: research design, review, and supervision; all authors conceptualized, analyzed the results and approved the manuscript.

Funding

The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

Supporting data are available from the Asan Medical Center, and due to ethical concerns and confidentiality agreements, data are available upon reasonable request. The data are available from the corresponding author Y.K. upon reasonable request.

Declarations

Ethics approval and consent to participate

This study obtained approval and waived the written informed consent from the Institutional Review Boards of Asan Medical Center (No. 2021–0321). All experiments were performed in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Division of Cardiology, Department of Internal Medicine, Asan Medical Center, University of Ulsan College of Medicine, 88, Olympicro 43Gil, Songpagu, Seoul 05505, Republic of Korea. [2]Department of Medical Science, Asan Medical Center, Asan Medical Institute of Convergence Science and Technology, University of Ulsan College of Medicine, 88, Olympicro 43 Gil, Sonpagu, 05505 Seoul, Republic of Korea. [3]Big Data Research Center, Asan Institute for Life Sciences, Asan Medical Center, 88, Olympicro 43Gil, Songpagu, Seoul 05505, Republic of Korea.

Received: 2 August 2023 Accepted: 7 November 2024
Published online: 20 November 2024

References

1. Cosgrove S. The relationship between antimicrobial resistance and patient outcomes: mortality, length of hospital stay, and health care costs. Clin Infect Dis. 2006;42:82–9.
2. Bauer M, Fitzgerald L, Haesler E, Manfrin M. Hospital discharge planning for frail older people and their family. Are we delivering best practice? A review of the evidence. J Clin Nurs. 2009;18:2539–46.
3. Tak J, Ruhnke W, Meltzer O. Association of patient preferences for participation in decision making with length of stay and costs among hospitalized patients. JAMA Intern Med. 2013;173:1195–205.
4. Kudyba S, Gregorio T. Identifying factors that impact patient length of stay metrics for healthcare providers with advanced analytics. Health Informatics. 2010;16:235–45.
5. Siddique M, et al. Interventions to reduce hospital length of stay in high-risk populations: a systematic review. JAMA Netw Open. 2021;4:e2125846.
6. Thornburgh Z, Samuel D. Factors influencing length of stay and discharge destination of patients with hip fracture rehabilitating in a private care setting. Geriatrics. 2022;7:44.
7. Iwase S, et al. Prediction algorithm for ICU mortality and length of stay using machine learning. Sci Rep. 2022;12:12912.
8. Wollny K, Pitt T, Brenner D, Metcalfe A. Predicting prolonged length of stay in hospitalized children with respiratory syncytial virus. Pediatr Res. 2022;92(6):1780–6.
9. Lior T, et al. Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. Expert Syst Appl. 2017;78:376–85.
10. Choi B, et al. Development of Predictive Model for Length of Stay(LOS) in Acute Stroke Patients using Artificial Intelligence. J Digital Convergence. 2018;16:231–42.
11. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. J Am Med Inform Assoc. 2016;23:2–10.
12. Cai X, et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. J Am Med Inform Assoc. 2016;23:553–61.
13. Haghparast-Bidgoli H, et al. Factors affecting hospital length of stay and hospital charges associated with road traffic-related injuries in Iran. BMC Health Serv Res. 2013;13:281.
14. Khosravizadeh O, et al. Factors affecting length of stay in teaching hospitals of a middle-income country. Electron Physician. 2016;5:3042–7.
15. Baek H, et al. Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. PLoS One. 2018;13:e0195901.
16. Odonkor C, et al. New utility for an old tool: can a simple gait speed test predict ambulatory surgical discharge outcomes? Am J Phys Med Rehabil. 2013;92:849–63.
17. Wang X, et al. Association between type of anesthesia and length of hospital stay in primary unilateral total knee arthroplasty patients: a single-center retrospective study. J Orthop Surg Res. 2021;16:671.
18. Bert F, et al. Predicting length of stay and discharge destination for surgical patients: a cohort study. Int J Environ Res Public Health. 2020;17:9490.
19. Rotar P. *et al.* Prediction of Prolonged Intensive Care Unit Length of Stay Following Cardiac Surgery. Elsevier. 2021.
20. Ong P, Pua Y. A prediction model for length of stay after total and unicompartmental knee replacement. Bone Joint J. 2013;95:1490–6.
21. Jo Y, et al. Prediction of Prolonged Length of Hospital Stay After Cancer Surgery Using Machine Learning on Electronic Health Records: Retrospective Cross-sectional Study. JMIR Med Inform. 2021;9:e23147.
22. Yang H, et al. Strategies for building robust prediction models using data unavailable at prediction time. J Am Med Inform Assoc. 2021;29:72–9.
23. Kim S. Factors influencing length of stay at the recovery room among elderly patients undergone general anesthesia. Korean J Adult Nurs. 2011;23(1):87–99.
24. Famili F, Shen W, Weber R, Simoudis E. Data Preprocessing and intelligent data analysis. Intell Data Anal. 1997;1:3–23.
25. Kotsiantis S, Kanellopoulos D, Pintelas P. Data Preprocessing for supervised learning. Int J Comput Inform Eng. 2007;1:4104–9.
26. Girish C, Ferat S. A survey on feature selection methods. Comput Electr Eng. 2014;40:16–28.
27. Arthur H, Robert K. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;1:55–67.
28. Tianqi C, Carlos G. Xgboost: A scalable tree boosting system. Int Conf Knowl Discov Data Mining. 2016;16:785–94.
29. Biau G, Scornet E. A random forest guided tour. TEST. 2016;25:197–227.
30. Imran K, Mevlut T, Turhan K. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl. 2008;34:366–74.
31. Oshiro T, Perez P, Baranauskas J. How Many Trees in a Random Forest? *MLDM* (2012).
32. Segal, M. Machine learning Benchmarks and random forest regression. Center forBioinform Mol Biostat. (2004).
33. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
34. Abbas, A. *et al*. Machine learning using preoperative patient factors can predict duration of surgery and length of stay for total knee arthroplasty. Elsevier. (2022).
35. Lundberg S, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67.
36. Lundberg S, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2:749–60.
37. Sutton R, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
38. Khalifa M. Clinical decision support: strategies for success. Procedia Comput Sci. 2014;37:422–7.
39. Kriegova E, et al. A theoretical model of health management using data-driven decision-making: the future of precision medicine and health. J Transl Med. 2021;19:68.
40. Silvério R, et al. Primary care physicians' decision-making processes in the context of multimorbidity: protocol of a systematic review and thematic synthesis of qualitative research. BMJ Open. 2019;9:e023832.
41. Shameer, K. et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai heart failure cohort. Pac Symp Biocomput. (2017).
42. Rath S, Rajaram K, Mahajan A. Integrated anesthesiologist and room scheduling for surgeries: Methodology and application. Oper Res. 2017;65(6):1460–78. https://doi.org/10.1287/opre.2017.1620.
43. Misic VV, Rajaram K, Gabel E. A simulation-based evaluation of machine learning models for clinical decision support: Application and analysis using hospital readmission. NPJ Digital Medicine. 2021;4(1):98. https://doi.org/10.1038/s41746-021-00461-7.
44. Bertsimas D, Pauphilet J, Stevens J, Tandon M. Length-of-stay and mortality prediction for a major hospital through interpretable machine learning. Manufacturing and Service Operations Management. (2020).
45. Bestsennyy O, Cordina J. The role of Personalization in the care journey: An example of patient engagement to reduce readmissions. McKinsey & Company. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-role-of-personalization-in-the-care-journey-an-example-of-patient-engagement-to-reduce-readmissions (2021).
46. Daghistani T, et al. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. Int J Cardiol. 2019;288:140–7.

47. Chua J, et al. Factors associated with prolonged length of stay in patients admitted with severe hypoglycaemia to a tertiary care hospital. Endocrinol Diabetes Metab. 2019;2:3.
48. Warnke I, Rössler W, Herwig U. Does psychopathology at admission predict the length of inpatient stay in psychiatry? Implications for financing psychiatric services. BMC Psychiatry. 2011;11:120.
49. Lin K, Lin H, Yeh P. Determinants of prolonged length of hospital stay in patients with severe acute ischemic stroke. J Clin Med. 2022;11:3457.
50. Lee S, et al. Factors associated with prolonged length of stay for elective hepatobiliary and neurosurgery patients: a retrospective medical record review. BMC Health Serv Res. 2018;18:5.
51. Eskandari M, et al. Evaluation of factors that influenced the length of hospital stay using data mining techniques. BMC Med Inform Decis Mak. 2022;22:280.
52. Naessens J, et al. Effect of illness severity and comorbidity on patient safety and adverse events. Am J Med Qual. 2012;27:48–57.
53. Moore L, et al. Impact of socio-economic status on hospital length of stay following injury: a multicenter cohort study. BMC Health Serv Res. 2015;15:285.
54. Hongbo H, et al. Is severity of family burden a correlate of length of stay? Psychiatry Res. 2015;230:84–9.

Publisher's Note