Computational and Structural Biotechnology Journal 20 (2022) 2861–2870

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Application of interpretable machine learning for early prediction of prognosis in acute kidney injury

Chang Hu [a,b,1], Qing Tan [c,1], Qinran Zhang [c,1], Yiming Li [a,b,1], Fengyun Wang [a,b], Xiufen Zou [c,*], Zhiyong Peng [a,b,*]

[a] Department of Critical Care Medicine, Zhongnan Hospital of Wuhan University, Wuhan, Hubei 430071, China
[b] Clinical Research Center of Hubei Critical Care Medicine, Wuhan, Hubei 430071, China
[c] School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China

ARTICLE INFO

Article history:
Received 27 February 2022
Received in revised form 31 May 2022
Accepted 1 June 2022
Available online 03 June 2022

Keywords:
Machine learning
Interpretability
Acute kidney injury
Critically illness
Mortality

ABSTRACT

Background: This study aimed to develop an algorithm using the explainable artificial intelligence (XAI) approaches for the early prediction of mortality in intensive care unit (ICU) patients with acute kidney injury (AKI).

Methods: This study gathered clinical data with AKI patients from the Medical Information Mart for Intensive Care IV (MIMIC-IV) in the US between 2008 and 2019. All the data were further randomly divided into a training cohort and a validation cohort. Seven machine learning methods were used to develop the models for assessing in-hospital mortality. The optimal model was selected based on its accuracy and area under the curve (AUC). The SHapley Additive exPlanation (SHAP) values and Local Interpretable Model-Agnostic Explanations (LIME) algorithm were utilized to interpret the optimal model.

Results: A total of 22,360 patients with AKI were finally enrolled in this study (median age, 69.5 years; female, 42.8%). They were randomly split into a training cohort (16770, 75%) and a validation cohort (5590, 25%). The eXtreme Gradient Boosting (XGBoost) model achieved the best performance with an AUC of 0.890. The SHAP values showed that Glasgow Coma Scale (GCS), blood urea nitrogen, cumulative urine output on Day 1 and age were the top 4 most important variables contributing to the XGBoost model. The LIME algorithm was used to explain the individualized predictions.

Conclusions: Machine-learning models based on clinical features were developed and validated with great performance for the early prediction of a high risk of death in patients with AKI.

1. Background

Acute kidney injury (AKI), recognized as a serious disorder that acutely affects the kidney's filtration function, is very common in critically ill patients and has a high mortality rate [1–3]. A recent multinational cross-sectional study reported that AKI occurred in more than half of patients treated in the intensive care unit (ICU) [4]. Meanwhile, increasing AKI severity was strongly associated with increased mortality even after adjusting for possible confounders [4]. Despite the large number of new therapeutic strategies that have been conducted, no effective treatment has consistently demonstrated clinical benefits.

Early identification of AKI patients at high risk for clinical deterioration is of great importance and may help to deliver proper care and optimize the use of limited resources. Considering the potential benefits of electronic alerts in AKI, many researchers have developed a variety of machine learning-based models to predict mortality for AKI patients [5–8]. However, these previously established models were limited to clinical implementation due to the minimal interpretability and black box nature of the algorithms [9]. Opening the black box is crucial because it can allow clinicians to easily understand the internal logic of machine learning (ML) [10,11]. Recently, Explainable Artificial Intelligence (XAI) has been introduced to address the fundamental question about the rationale of the decision-making process in ML. The most widespread explainable techniques comprise SHapley Additive exPlanation

* Corresponding authors at: Department of Critical Care Medicine, Zhongnan Hospital of Wuhan University, Wuhan, Hubei 430071, China (Z. Peng).
E-mail addresses: xfzou@whu.edu.cn (X. Zou), Pengzy5@hotmail.com (Z. Peng).
[1] Co-first authors (these authors have contributed equally).

https://doi.org/10.1016/j.csbj.2022.06.003
2001-0370/

(SHAP) [12] and Local Interpretable Model-Agnostic Explanations (LIME). These new interpretable methods have been successfully applied to explain the ML models related to mortality prediction in acute gastrointestinal bleeding [13] and sepsis [14], prediction of antimicrobial resistance [15] and the occurrence of AKI following cardiac surgery [16].

However, as far as we know, there has been relatively little analysis of the reliability and robustness of the explanation methods in outcome prediction among AKI. Therefore, in this study, we aimed to use a ML approach to predict in-hospital mortality in critically ill patients with AKI and utilize XAIs to increase the interpretability, fairness, and transparency of ML.

2. Methods

2.1. Ethics

The establishment of the Medical Information Mart for Intensive Care IV (MIMIC-IV version 1.0) database was approved by the Massachusetts Institute of Technology (Cambridge, MA) and Beth Israel Deaconess Medical Center (Boston, MA), and patients provided their consent to have their data captured in the database. Thus, the ethical approval statement was waived in this study, as the data in the MMIC-IV database were unidentifiable.

2.2. Data source

This study gathered clinical data from the MIMIC-IV database between 2008 and 2019. MIMIC-IV is a large, single-center, open-access database comprising 76,540 ICU admissions [17]. We accessed the MIMIC-IV after completion of the Protecting Human Research Participants exam (Record ID: 47460147). This study was conducted in accordance with the principles of the Declaration of Helsinki in 2013, and all reporting followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) Statement [18].

2.3. Study population

All adult (aged 18 years old and older) individuals diagnosed with AKI were enrolled in this study. AKI was defined and classified according to the Kidney Disease: Improving Global Outcomes (KDIGO) criteria during the first 24 h after ICU admission [19]. In brief, the KDIGO criteria included an increase in serum creatinine by $\geq$ 0.3 mg/dl within 48 h; an increase in serum creatinine to $\geq$ 1.5 times baseline, which is known or presumed to have occurred within the prior 7 days; or urine volume <0.5 mL/kg/hour for 6 h. Details on the definition and classification of AKI are provided in Supplementary Table S1. For patients with multiple ICU admissions during hospitalization, only the first admission was included for analysis. We excluded patients with an ICU length of stay of <3 h.

2.4. Data collection and preprocessing of data

We used structured query language (SQL) programming in Navicat Premium (version 15) to extract clinical data from the MIMIC-IV database. The information collected from the database followed the Deshmukh et al. [13] procedure. We collected information related to the patients' demographic characteristics, history of chronic diseases, vital signs, laboratory findings, medical treatments, severity scores of illness and outcomes.

Demographic variables collected for the study included age, sex, body weight and height. Medical conditions included hypertension, diabetes, congestive heart failure, cerebrovascular disease, chronic pulmonary disease, liver disease, tumor and acquired immune deficiency syndrome. We collected mean values in the first 24 h after ICU admission for the vital sign data, including heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, respiratory rate, body temperature and SpO_2. For laboratory findings, we collected the maximum value in the first 24 h after ICU admission including blood glucose, lactate, pH, pCO_2, base excess, white blood cell count, anion gap, bicarbonate, blood urea nitrogen, serum calcium, serum chloride, serum creatinine, serum sodium, serum potassium, serum fibrinogen, international normalized ratio, prothrombin time, partial thromboplastin time, alanine aminotransferase, alkaline phosphatase, aspartate aminotransferase, total bilirubin, creative phosphokinase, creatine kinase MB and lactate dehydrogenase. The minimum values in the first 24 h after ICU admission were selected for the PaO_2/FiO_2 ratio, hematocrit, hemoglobin, platelets, albumin and globulin. Medical treatments included the use of antibiotics, mechanical ventilation and vasopressors during the first 24 h after ICU admission. For the severity scores of illness, we calculated the maximum value for Sequential Organ Failure Assessment (SOFA), Oxford Acute Severity of Illness Score (OASIS), Simplified Acute Physiology Score II (SAPS-II), and the minimum value for Glasgow Coma Scale (GCS) during the first 24 h after ICU admission. We also collected the cumulative urine output during the first 24 h after ICU admission. A summary of each variable can be found in Supplementary Table S2. The code is available at https://github.com/MIT-LCP/mimic-iv.

Variables with more than 20% missing values were removed from further analysis. The multiple imputation method, recognized as a better approach to deal with missing observations in both outcome and independent variables, was used to handle missing data below 20% using the 'mice' package in R. To avoid overfitting, we used least absolute shrinkage and selection operator (LASSO) regression to identify potential variables associated with mortality.

2.5. Statistical analysis

The normal distribution of continuous variables was determined by the Kolmogorov-Smirnov test. The normally distributed variables were described as the means ± standard deviations (SD), while the skewed distributed variables were expressed as the median and interquartile range (IQR), and the categorical variables were expressed as number and percentages. Continuous variables between groups were compared by Student's *t*-test or the Mann–Whitney *U* test, as appropriate. Categorical variables between groups were compared by Pearson's chi-squared test or Fisher's exact test, as appropriate.

All statistical analyses were performed using Python (Version 3.6.6) and R software (Version 3.6.1, R Foundation for Statistical Computing). A *P* value (2-sided) below 0.05 was considered as statistically significant.

2.6. Machine learning model for feature mining and feature visualization

Stable and significant features were very important to predict the risks of in-hospital mortality, and the feature-related in-hospital risks were studied with feature mining and feature visualization. In feature mining, all patients enrolled in the study were randomly split into a training set (75%) and a validation set (25%). We used the LASSO [20] algorithm to reduce the dimensionality of the features. The seven ML methods [Support Vector Machine (radial bias function) (SVM)[21], k-Nearest Neighbors (KNN) [22], eXtreme Gradient Boosting (XGBoost) [12], Decision Tree (DT)[23], Naive Bayes (NB) [24], Random Forest (RF) [25] and logistic regression (LR) [26] were used to develop and validate

the models for assessing risks of in-hospital mortality. Predictive performance for the classifiers was evaluated using the accuracy, area under the receiving operating characteristic curve (AUC), sensitivity, and specificity measures. DeLong's test was performed to assess the differences in AUC. The calibration curve was used to compare the prediction probability of the models and the ground truth. Our final candidate model was selected based on the accuracy and the AUC. First, we used the SHAP values to visualize the significant features that influence the risks of mortality, to analyze the importance of individual features affecting the output of the model and to visualize the impact of key features on the final model in individuals. Next, we conducted the LIME [27] algorithm to fit the predictive behavior of the model. The details of the procedure in this study are shown in Fig. 1. Subgroup analyses based on different KDIGO stages were performed. Sensitivity analyses were conducted for the outcome by removing the most crucial factor from our final candidate model.

3. Results

3.1. Participants

A total of 33,020 participants diagnosed with AKI were screened for eligibility; of these 33,020 patients, 10,657 were excluded due to multiple ICU admission (only the first admission was included

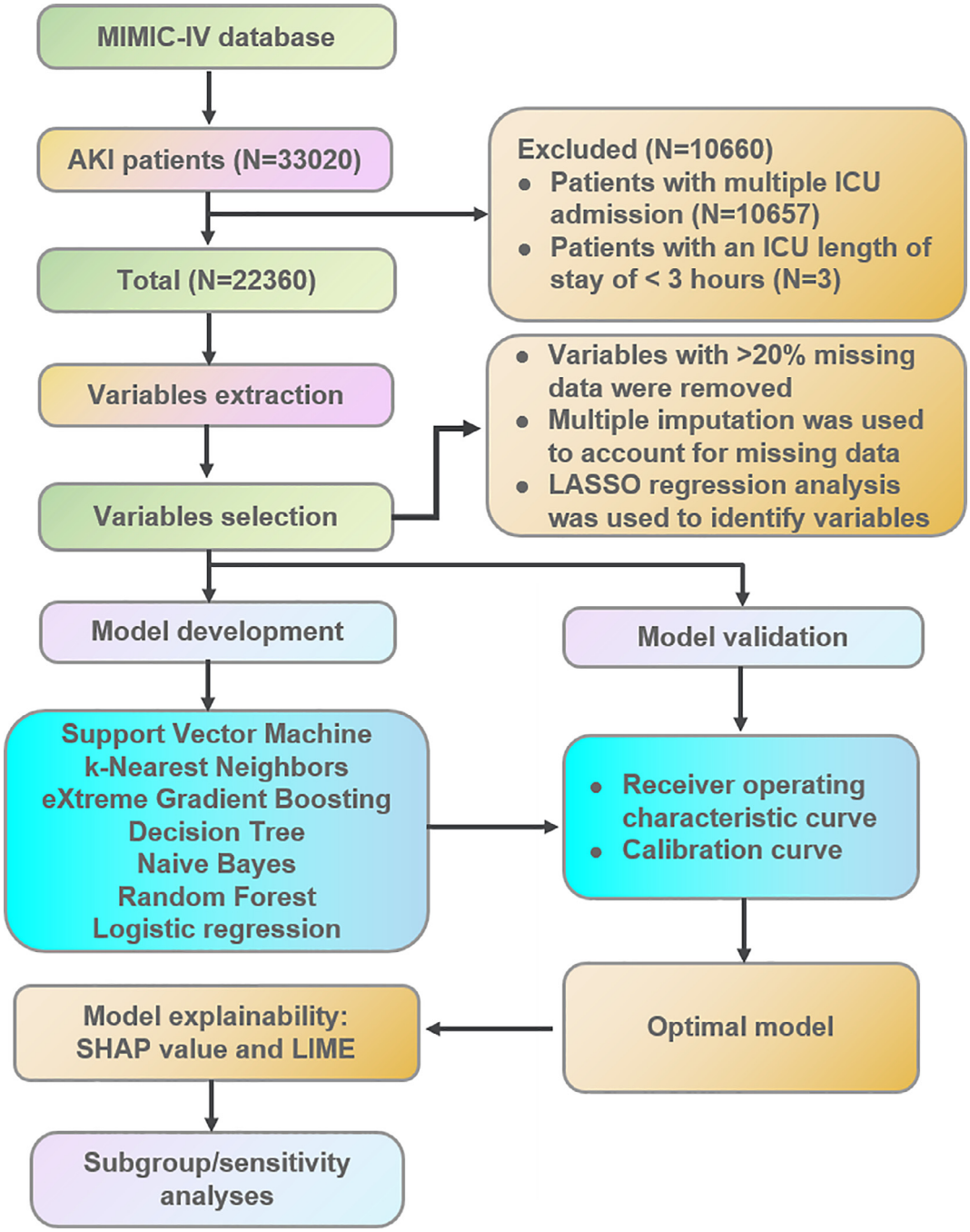


Fig. 1. Flowchart of this study. MIMIC-IV, Medical Information Mart for Intensive Care IV; LASSO, least absolute shrinkage and selection operator; SHAP, SHapley Additive explanation; LIME, Local Interpretable Model-Agnostic Explanations.

for analysis), and 3 were excluded due to an ICU length of stay of <3 h. Finally, 22,360 patients were eligible for participation (Fig. 1). The prevalence of in-hospital mortality was 15.6% (3484/22360). These patients had a median age of 69.5 (IQR, 58.3–79.9) years and 42.8% (9574/22360) were female. Hypertension (9245/22360, 41.3%), diabetes (7268/22360, 32.5%) and congestive heart failure (6819/22360, 30.5%) ranked as the top 3 comorbidities. The baseline characteristics of the dataset are summarized in Table 1.

3.2. Predictor selection

Nineteen variables with more than 20% missing values were removed (Supplementary Fig. S1). Multiple imputation by chained equation was used to impute missing values for each variable below 20%. Forty variables measured on the first day after ICU admission were included in the LASSO regression. After LASSO regression selection with 5-fold cross-validation via minimum criteria, 29 variables remained as significant predictors for mortality (Supplementary Fig. S2), and the details of the variables are shown in Supplementary Table S3.

3.3. Model development and validation

A total of 22,360 patients were selected and randomly split into a training cohort (16770, 75%) and a validation cohort (5590, 25%). We used 29 variables selected by LASSO regression as input factors (Supplementary Table S3), and seven ML methods, including SVM, KNN, XGBoost, DT, NB RF and LR, were established to predict outcome. In the validation cohort, the XGBoost model achieved the

Table 1
Demographic and clinical characteristics at baseline.

Characteristics	Total (n = 22360)	Survivors (n = 18876)	Non-survivors (n = 3484)	*P* value
Demographics				
Age, year	69.5(58.3–79.9)	68.8(57.8–79.2)	73.4(61.4–83.4)	<0.001
Sex				0.001
Male, n (%)	12786(57.2)	10886(57.7)	1900(54.5)	
Female, n (%)	9574(42.8)	7990(42.3)	1584(45.5)	
Weight, kg	83(70–99)	84(70–100)	78(65–94)	<0.001
Comorbidities, n (%)				
Hypertension	9245(41.3)	7915(41.9)	1330(38.2)	<0.001
Diabetes	7268(32.5)	6218(32.9)	1050(30.1)	0.001
Congestive heart failure	6819(30.5)	5631(29.8)	1188(34.1)	<0.001
Cerebrovascular disease	3215(14.4)	2505(13.3)	710(20.4)	<0.001
Chronic pulmonary disease	5694(25.5)	4779(25.3)	915(26.3)	0.239
Liver disease	3062(13.7)	2129(11.3)	933(26.8)	<0.001
Renal disease	5328(23.8)	4361(23.1)	967(27.8)	<0.001
Tumor	3259(14.6)	2431(12.9)	828(23.8)	<0.001
Acquired immune deficiency syndrome	86(0.4)	69(0.4)	17(0.5)	0.284
Vital signs on Day 1				
Heart rate, bpm	84(74–95)	83(74–93)	91(77–104)	<0.001
Systolic blood pressure, mmHg	114(106–125)	115(107–125)	108(100–120)	<0.001
Diastolic blood pressure, mmHg	60(54–67)	60(54–67)	59(52–66)	<0.001
Mean arterial pressure, mmHg	75(70–82)	76(70–83)	73(67–80)	<0.001
Respiratory rate	19(17–21)	18(16–21)	21(18–24)	<0.001
Body temperature, ℃	36.8(36.6–37.1)	36.8(36.6–37.1)	36.7(36.4–37.1)	<0.001
SpO_2, %	93(90–95)	93(91–95)	91(86–94)	<0.001
Laboratory findings on Day 1				
White blood cell, K/uL	13.6(9.9–18.4)	13.3(9.8–17.8)	15.3(10.4–21.4)	<0.001
Hematocrit, %	30(25–34)	30(25–34)	29(24–35)	<0.001
Hemoglobin, g/dL	9.8(8.4–11.4)	9.9(8.4–11.5)	9.4(7.9–11.2)	<0.001
Platelets, K/uL	164(115–224)	165(118–224)	152(89–225)	<0.001
Blood urea nitrogen, mg/dL	23(16–37)	21(15–33)	35(22–54)	<0.001
Serum creatinine, mg/dL	1.2(0.9–1.9)	1.1(0.8–1.7)	1.8(1.1–2.8)	<0.001
International normalized ratio	1.3(1.2–1.6)	1.3(1.1–1.5)	1.6(1.2–2.3)	<0.001
Prothrombin time, s	14.6(12.8–17.5)	14.4(12.7–16.8)	17.0(13.7–24.9)	<0.001
Partial thromboplastin time, s	32.8(28.3–45.0)	32.1(28.0–41.9)	40.0(30.6–66.5)	<0.001
Blood glucose, mg/dL	169(135–214)	168(135–209)	178(138–248)	<0.001
Anion gap, mEq/L	16(13–19)	15(13–18)	19(16–24)	<0.001
Bicarbonate, mmol/L	24(22–26)	24(22–27)	22(19–25)	<0.001
Serum sodium, mEq/L	140(137–142)	140(137–142)	140(136–144)	<0.001
Serum potassium, mEq/L	4.5(4.1–5.0)	4.5(4.1–5.0)	4.7(4.2–5.4)	<0.001
Serum calcium, mg/dL	8.5(8.0–9.0)	8.5(8.0–9.0)	8.5(8.0–9.1)	0.004
Serum chloride, mEq/L	106(102–110)	106(103–110)	106(101–110)	<0.001
Medical treatments, n (%)				
Antibiotics	9281(41.5)	7356(39.0)	1925(55.3)	<0.001
Mechanical ventilation	10508(47.0)	8627(45.7)	1881(54.0)	<0.001
Vasopressors	6398(28.6)	4855(25.7)	1543(44.3)	<0.001
Urine output on Day 1, mL	1180(735–1760)	1255(820–1820)	727(300–1286)	<0.001
Severity scores of illness				
GCS	14(11–15)	14(12–15)	11(5–15)	<0.001
SOFA	5(3–8)	5(3–7)	10(6–13)	<0.001
OASIS	34(28–41)	33(27–39)	43(37–49)	<0.001
SAPS-II	38(30–48)	36(29–45)	53(42–65)	<0.001

Data were reported as no. (%) or median (IQR).
GCS, Glasgow Coma Scale; SOFA, Sequential Organ Failure Assessment; OASIS, Oxford Acute Severity of Illness Score; SAPS-II, Simplified Acute Physiology Score II.

best performance with an AUC of 0.890 (RF: 0.885; NB: 0.842; LR: 0.840; SVM: 0.756; KNN: 0.674; DT: 0.676; respectively) (Fig. 2A). To further evaluate the performance of the seven models, precision, sensitivity and specificity were also calculated and the results are shown in Table 2. Although the AUC in the XGBoost model was numerically higher than that the RF model (0.890 vs. 0.885) without achieving statistical significance (P = 0.342) (Supplementary Table S4), the visual inspection of the calibration plots suggested that the XGBoost model was superior to the RF model in consistency (Supplementary Fig. S3). Similarly, the XGBoost model performance surpassed those of other clinical scores for disease severity [SAPS-II (AUC): 0.782; OASIS (AUC): 0.782] (Fig. 2B). Thus, the XGBoost model was selected for further prediction.

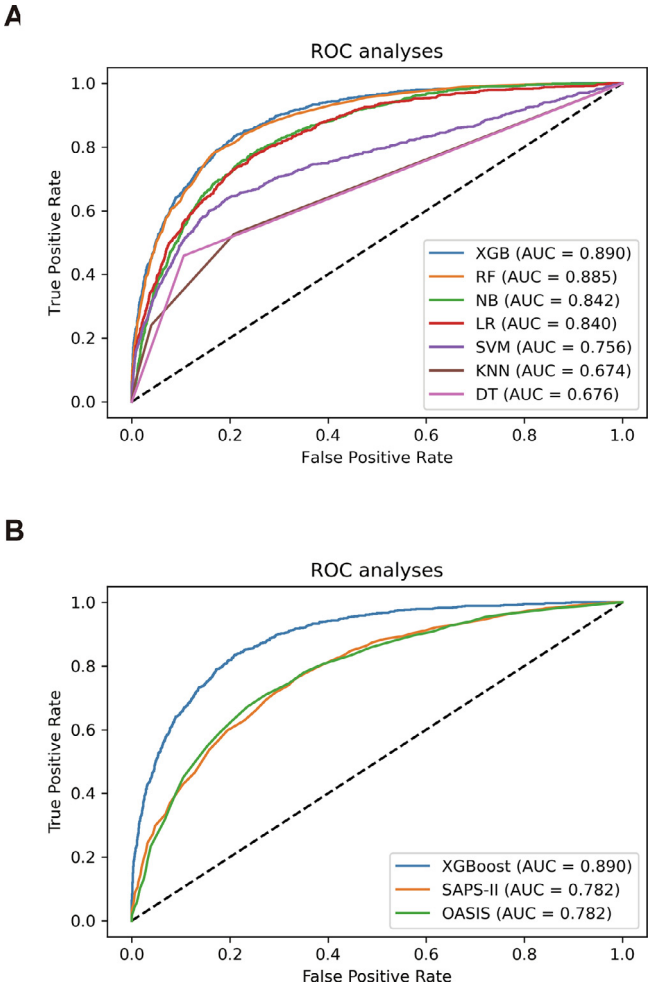


Fig. 2. ROC curves for the machine-learning models and traditional severity of illness scores to predict in-hospital mortality. (A) AUCs are shown for all 7 machine-learning models in the validation cohort. (B) Comparison of the AUC value between the XGBoost model and traditional severity of illness scores invalidation cohort (OASIS and SAPS-II). XGB, eXtreme Gradient Boosting; RF, Random Forest; SVM, Support Vector Machine (radial bias function); LR, Logistic Regression; NB, Naive Bayes; KNN, k-Nearest Neighbors; DT, Decision Tree; OASIS, Oxford Acute Severity of Illness Score; SAPS-II, Simplified Acute Physiology Score II; AUC, the area under curve; ROC, receiving operating characteristic curve.

3.4. Model explainability

We tried to open the 'black box' in the XGBoost model by SHAP values and explained how the model works in predicting mortality. The feature importance ranking with the SHAP summary plot for the XGBoost model is presented in Fig. 3A, and the top 4 most important variables contributing to the model were GCS, blood urea nitrogen, cumulative urine output on Day 1, and age. Additionally, we depicted how a single variable affected the output of the XGBoost prediction model using SHAP dependence analysis (Fig. 3B). More detailed results of the top 4 most important clinical features affecting the output of the XGBoost prediction model are shown in Fig. 4.

Then, we used SHAP force analysis and the LIME algorithm to explain the individualized prediction of death by drawing two samples from the validation set. Fig. 5A and Fig. 5B present a deceased case using SHAP force analysis and the LIME algorithm, respectively. This case was an 81-year-old woman with a history of congestive heart failure and chronic pulmonary disease admitted to the ICU for AKI. The predicted probability for death by the XGBoost model was 93%. Factors detected by the XGBoost model that increased the risk of death were GCS score of 3, urine output on Day 1 of 162 mL, blood urea nitrogen of 45 mg/dL and serum sodium of 154 mmol/L. Factors that decreased the risk of death included the lack of a history of cerebrovascular disease or tumor. The outcome predicted by the XGBoost model was death for this patient and the actual outcome was also death. Similarly, Fig. 5C and Fig. 5D present a survival case using SHAP force analysis and the LIME algorithm, respectively. This case was a 79-year-old woman admitted to the ICU for AKI. The predicted probability for death by the XGBoost model was 17%. The patient's elevated serum sodium of 150 mEq/L, a need for mechanical ventilation on Day 1 and systolic blood pressure of 102 mmHg contribute to increasing the death risk, while a GCS score of 14, the cumulative urine output of 1790 mL on Day 1 and no history of tumor or cerebrovascular disease contributed to decreasing the patient's death risk. The outcome predicted by the XGBoost model was survival for this patient, and the actual outcome was also survival.

3.5. Subgroup analyses

Prespecified subgroup analyses conducted for the different KDIGO stages showed that the XGBoost model remained robust in predicting mortality among patients with KDIGO stage 1

Table 2
Performance of the seven ML models for predicting in-hospital mortality.

ML models	Accuracy, %	AUC, 95% CI	Sensitivity, %	Specificity, %
XGBoost	87.7	0.890(0.880–0.897)	82.5	79.6
RF	87.5	0.885(0.876–0.893)	78.2	83.2
NB	84.3	0.842(0.833–0.852)	74.1	78.6
LR	86.0	0.840(0.830–0.849)	77.3	75.2
SVM	83.8	0.756(0.742–0.765)	65.0	78.4
KNN	84.4	0.674(0.662–0.687)	52.7	79.1
DT	82.6	0.676(0.665–0.683)	44.4	89.4

ML, machine learning; XGBoost, eXtreme Gradient Boosting; RF, Random Forest; SVM, Support Vector Machine (radial bias function); LR, Logistic Regression; NB, Naive Bayes; KNN, k-Nearest Neighbors; DT, Decision Tree; AUC, the area under curve.

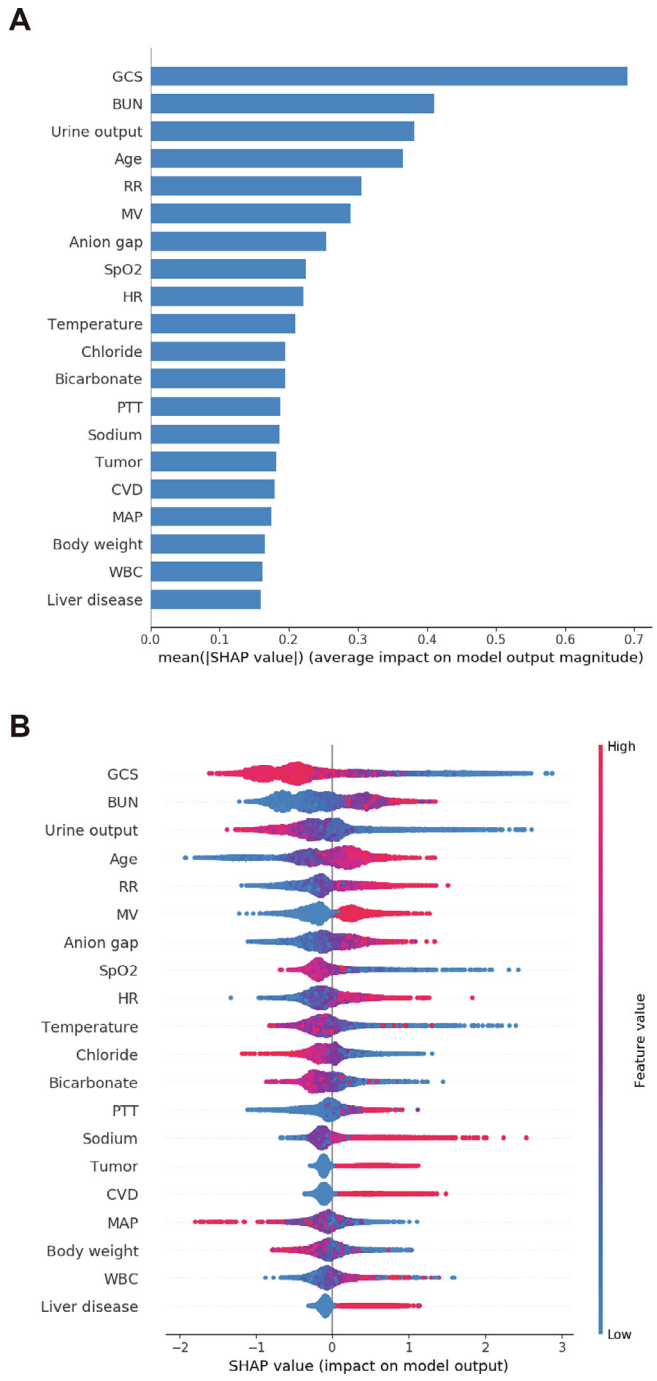


Fig. 3. SHAP summary plot for the top 20 clinical features contributing to the XGBoost model. (A) Ranking of feature importance indicated by SHAP. The matrix plot depicts the importance of each covariate in the development of the final predictive model. (B) The attributes of the features in the black box model. Each line represents a feature, and the abscissa is the SHAP value. Red dots represent higher feature values, and blue dots represent lower feature values. SHAP, SHapley Additive explanation; XGBoost, eXtreme Gradient Boosting. GCS, Glasgow Coma Scale; BUN, blood urea nitrogen; RR, respiratory rate; MV, mechanical ventilation; HR, heart rate; PTT, partial thromboplastin time; CVD, cerebrovascular disease; MAP, mean arterial pressure; WBC, white blood cell. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(N = 1184), KDIGO stage 2 (N = 2823) and KDIGO stage 3 (N = 1583) (AUC: 0.929; 0.888; 0.830; respectively). The detailed results are shown in Supplementary Fig. S4.

3.6. Sensitivity analyses

When excluding the GCS score in the XGBoost model, the AUC for mortality prediction in AKI patients was 0.830 (Supplementary Fig. S5). In addition, the prediction results of the XGBoost model on samples were counted in the validation set, and the distribution of the two cases (one survival and one deceased) of the aggregated prediction results was also visualized using the LIME algorithm (Supplementary Fig. S6).

4. Discussion

In this study, we developed and validated seven ML methods using twenty-nine clinical variables to assess the risks of in-hospital mortality in critically ill patients with AKI. The XGBoost model exhibited the best performance in terms of discrimination and accuracy. Subgroup analyses also revealed that the XGBoost model achieved robust performance for the prediction of death in different KDIGO stages. Additionally, SHAP values were used to reveal the feature importance and how particular compound substructures influence the XGBoost prediction. GCS, blood urea nitrogen, cumulative urine output on Day 1 and age were the top 4 most important variables contributing to the XGBoost model. Finally, the LIME algorithm was used to facilitate the individualized predictions.

Accurate and timely prediction of mortality for AKI is required to identify patients at high risk of clinical deterioration so that preventive measures can be taken in a timely manner, which may reduce mortality. Several studies have attempted to establish prognostic models among AKI patients with ML methods and showed a modest prognostic yield [5,7,8,28–34]. For example, a study from the US constructed a prognostic model for predicting 60-day mortality in critically ill patients with AKI. They found that the predictive model with logistic regression yielded an AUC of 0.85 (95% CI: 0.83–0.88), surpassing those of APACHE II and SOFA [5]. However, this study did not perform internal or external validation. In 2019, Lin and colleagues constructed prediction models of mortality risk based on the RF, ANN (artificial neural network) and SVM for AKI patients. In the testing set, they found that the RF model had the largest AUC (0.866, 95% CI: 0.862–0.870) [7]. A recent modeling study for AKI also found that the XGBoost model achieved the best performance with an AUC of 0.796, compared to the LR, SVM and RF models (AUC: 0.662, 0.667, and 0.692, respectively) [28]. However, all the above ML models were established based on limited algorithm tools and were at a loss to adequately explain how they work. In the current study, we developed a variety of ML approaches containing the linear model (e.g., LR), kernel-based method (e.g., SVM), gradient boosting classifier (e.g., XGBoost) and other ML models (e.g., KNN, NB, DT and RF) and selected the optimal one with the best performance in discrimination and accuracy. Furthermore, we summarized the previous clinical prediction model related to mortality risk prediction in AKI patients and found that the XGBoost model in the current study performed best with an AUC of 0.890 in the validation cohort (Supplementary Table S5). Moreover, given the opaque black box nature of ML, we utilized the SHAP values and LIME algorithm to interpret and reveal the most important factors contributing to the prediction results, which greatly improved the model's interpretability.

In the current study, a summary of feature importance in the XGBoost model showed that the GCS score was the most crucial factor in the development of death in patients with AKI. The GCS score, as a common indicator of consciousness ranging from 3 to 15, can obviously show the neurological changes among critically ill patients [35]. Concurrently, there was a strong correlation between the GCS score and clinical outcome [36]. Nevertheless,

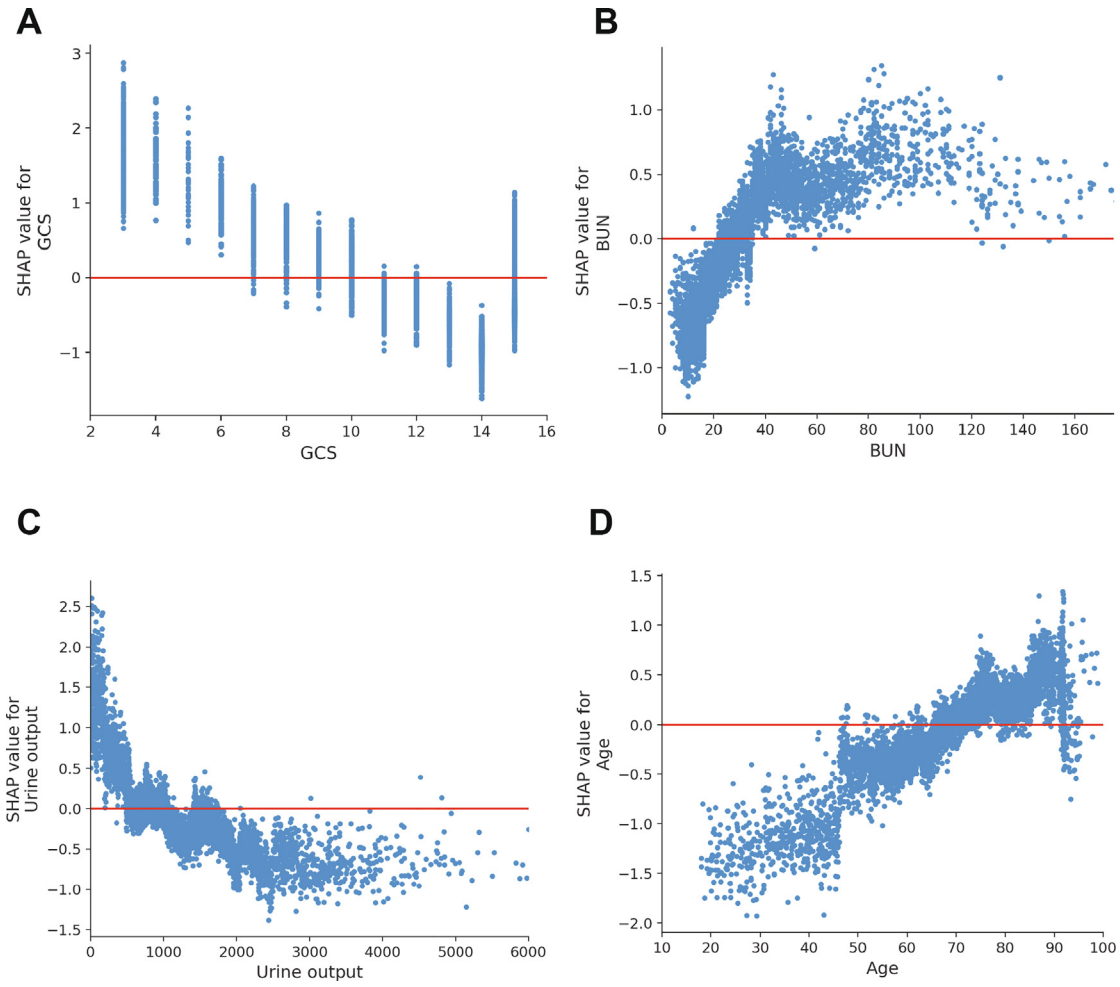
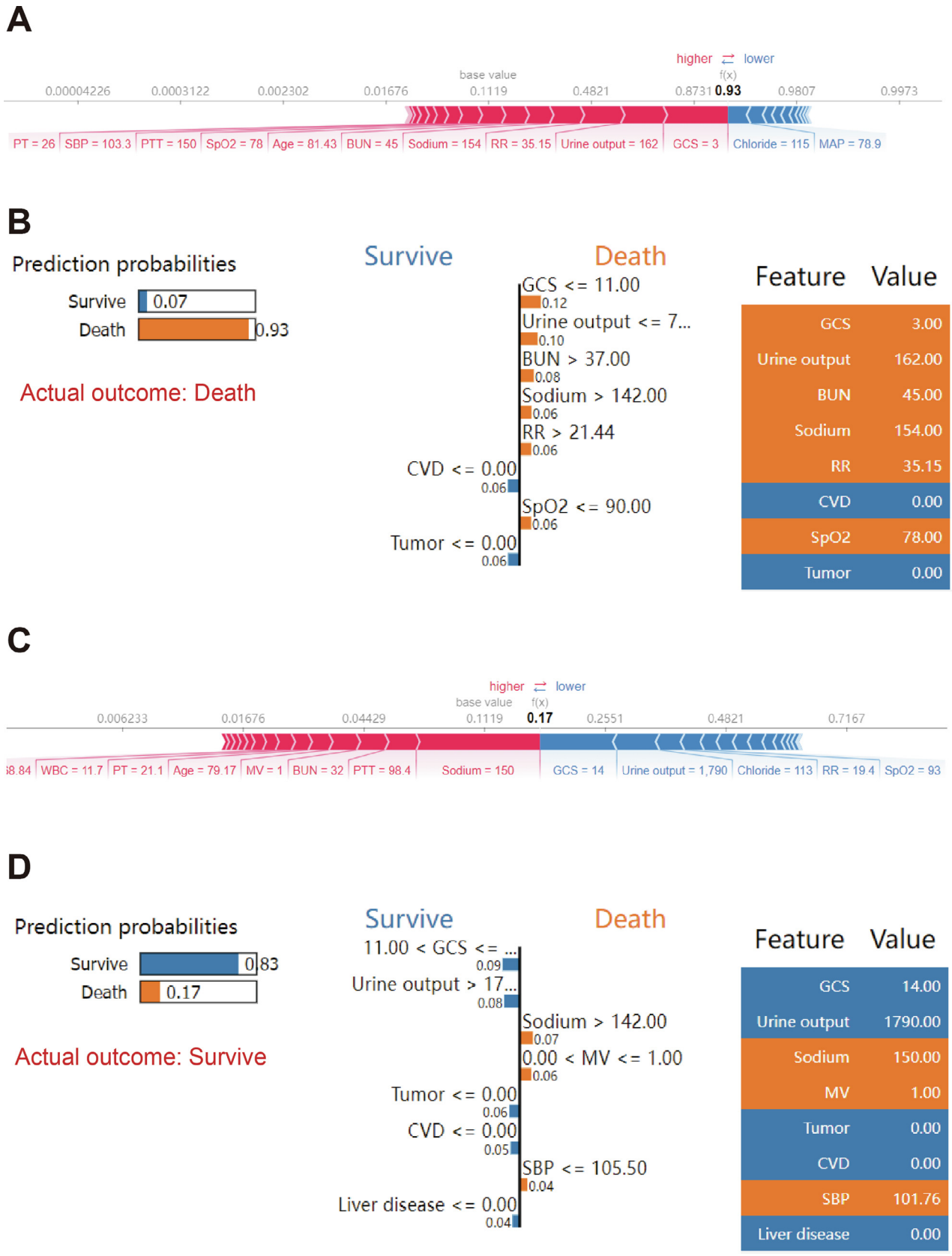


Fig. 4. SHAP dependence plot for the top 4 clinical features contributing to the XGBoost model. GCS; (B) BUN; (C) Urine output; (D) Age. SHAP values for specific features exceed zero, representing an increased risk of death. GCS, Glasgow Coma Scale; BUN, blood urea nitrogen; SHAP, SHapley Additive explanation; XGBoost, eXtreme Gradient Boosting.

none of the previous models included this critical feature in predicting mortality risk for AKI patients [5,7,8,28,29]. Our study found that blood urea nitrogen (ranked 2nd) and the cumulative urine output on Day 1 (ranked 3rd) were closely associated with mortality in AKI. Blood urea is excreted mainly by the kidney. The elevated blood urea nitrogen level indicated renal lesions, which also increased the risk of adverse outcomes [4]. Additionally, urine output has been found to be a marker for AKI and to be associated with adverse outcomes in critically ill patients [37]. Age was also another key factor for predicting mortality, which had been studied in previous reports as an independent predictor for hospital mortality among AKI patients[8].

The present study developed and validated a good performance model for mortality prediction in critically ill patients admitted for AKI, with a large sample size from the MIMIC-IV database and a precise definition of AKI. The model was built based on 29 candidate variables related to patients' demographic characteristics, medical history, vital signs, laboratory findings, medical treatments and GCS score, which was more systematic and robust than previously reported models that included only parts of investigative modalities. Overfitting is one of the critical problems in developing models by ML, which may lead to inadequate conclusions. In this study, we performed LASSO regression to reduce overfitting in the process of feature selection. This method surpasses the method of selecting predictors according to the strength of their univariable association with outcome. In addition, no other severity score of illness except GCS was enrolled as input variables for model construction due to the possible reuse of variables (such as SOFA score contains total bilirubin, vasopressor use, GCS, serum creatinine and urine output). Additionally, we attempted to use the latest methods (SHAP value and LIME algorithm) to interpret and analyze why and how the XGBoost model worked during execution. Although LIME and SHAP values showed a similar trend overall for model interpretation, they were also presented with distinguishing specific regions of slight mismatch. These differences indicated where features go from favoring survived to favoring deceased. Moreover, this early warning system is being translated for clinical implementation in cooperation with technology companies, and our team will look forward to its clinical use in the future.

However, there were also several limitations in this study. First, many laboratory parameters were removed before model construction due to missing data of over 20% (e.g., D-dimer, globulin, and albumin). Second, as a critical parameter to understand the pathological mechanisms of disease, the etiology of AKI was not recorded in the MIMIC-IV database, and we failed to add this factor to the model construction. These may have inevitably caused selection bias. Third, the clinical indicators were gathered using a single-center database due to the nature of MIMIC-IV. Fourth, although XGBoost does take into account interactions between fea-


A
higher ⇄ lower
base value
f(x)
0.00004226
0.0003122
0.002302
0.01676
0.1119
0.4821
0.8731
0.93
0.9807
0.9973
PT = 26
SBP = 103.3
PTT = 150
SpO2 = 78
Age = 81.43
BUN = 45
Sodium = 154
RR = 35.15
Urine output = 162
GCS = 3
Chloride = 115
MAP = 78.9
B
Prediction probabilities
Survive 0.07
Death 0.93
Actual outcome: Death
Survive
Death
GCS <= 11.00
0.12
Urine output <= 7...
0.10
BUN > 37.00
0.08
Sodium > 142.00
0.06
RR > 21.44
0.06
CVD <= 0.00
0.06
SpO2 <= 90.00
0.06
Tumor <= 0.00
0.06
Feature Value
GCS 3.00
Urine output 162.00
BUN 45.00
Sodium 154.00
RR 35.15
CVD 0.00
SpO2 78.00
Tumor 0.00
C
higher ⇄ lower
base value
f(x)
0.006233
0.01676
0.04429
0.1119
0.17
0.2551
0.4821
0.7167
8.84
WBC = 11.7
PT = 21.1
Age = 79.17
MV = 1
BUN = 32
PTT = 98.4
Sodium = 150
GCS = 14
Urine output = 1,790
Chloride = 113
RR = 19.4
SpO2 = 93
D
Prediction probabilities
Survive 0.83
Death 0.17
Actual outcome: Survive
Survive
Death
11.00 < GCS <= ...
0.09
Urine output > 17...
0.08
Sodium > 142.00
0.07
0.00 < MV <= 1.00
0.06
Tumor <= 0.00
0.06
CVD <= 0.00
0.05
SBP <= 105.50
0.04
Liver disease <= 0.00
0.04
Feature Value
GCS 14.00
Urine output 1790.00
Sodium 150.00
MV 1.00
Tumor 0.00
CVD 0.00
SBP 101.76
Liver disease 0.00

tures, these potentially relevant interactions will not be shown by LIME as it is a linear model. In addition, although LIME provides an explanation, this explanation depends on the parameters used to develop the local model, and by changing the parameters, other (potentially very different) explanations can arise. Fifth, the models were validated internally only in this study and further multicenter external validation is needed to verify the model's discriminating ability and generalizability.

5. Conclusions

The ML models based on clinical features were developed and validated with great performance in the early prediction of a high risk of death in AKI. Application of the SHAP values and LIME algorithm in ML may help physicians in clinical decision-making.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the National Natural Science Foundation of China (grants 81772046 and 81971816 to Dr. Peng and 11831015 to Dr. Zou) and the Special Project for Significant New Drug Research and Development in the Major National Science and Technology Projects of China (2020ZX09201007 to Dr. Peng). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.003.

References

[1] Chawla LS, Amdur RL, Shaw AD, Faselis C, Palant CE, Kimmel PL. Association between AKI and long-term renal and cardiovascular outcomes in United States veterans. Clin J Am Soc Nephrol 2014;9(3):448–56.
[2] Bouchard J, Soroko SB, Chertow GM, et al. Fluid accumulation, survival and recovery of kidney function in critically ill patients with acute kidney injury. Kidney Int 2009;76(4):422–7.
[3] Coca SG, Yusuf B, Shlipak MG, Garg AX, Parikh CR. Long-term risk of mortality and other adverse outcomes after acute kidney injury: a systematic review and meta-analysis. Am J Kidney Dis 2009;53(6):961–73.
[4] Hoste EA, Bagshaw SM, Bellomo R, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. Intensive Care Med 2015;41(8):1411–23.
[5] Demirjian S, Chertow GM, Zhang JH, et al. Model to predict mortality in critically ill adults with acute kidney injury. Clin J Am Soc Nephrol 2011;6(9):2114–20.
[6] Zhou XD, Chen QF, Sun DQ, et al. Remodeling the model for end-stage liver disease for predicting mortality risk in critically ill patients with cirrhosis and acute kidney injury. Hepatol Commun 2017;1(8):748–56.
[7] Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. Int J Med Inform 2019;125:55–61.
[8] Huang H, Liu Y, Wu M, Gao Y, Yu X. Development and validation of a risk stratification model for predicting the mortality of acute kidney injury in critical care patients. Ann Transl Med 2021;9(4):323.
[9] Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886.
[10] . Respir Med 2018;6(11):801..
[11] Azodi CB, Tang J, Shiu SH. Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends Genet 2020;36(6):442–55.
[12] Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell 2020;2(1):56–67.
[13] Deshmukh F, Merchant SS. Explainable Machine Learning Model for Predicting GI Bleed Mortality in the Intensive Care Unit. Am J Gastroenterol 2020;115(10):1657–68.
[14] Hu C, Li L, Huang W, et al. Interpretable Machine Learning for Early Prediction of Prognosis in Sepsis: A Discovery and Validation Study. Infect Dis Ther 2022;11(3):1117–32.
[15] Weis C, Cuenod A, Rieck B, et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nat Med 2022;28(1):164–74.
[16] Tseng PY, Chen YT, Wang CH, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 2020;24(1):478.
[17] Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
[18] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594.
[19] Stevens PE, Levin A. Kidney Disease: Improving Global Outcomes Chronic Kidney Disease Guideline Development Work Group M. Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline. Ann Intern Med 2013;158(11):825–30.
[20] Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met 1996;58(1):267–88.
[21] Burges CJC. A tutorial on Support Vector Machines for pattern recognition. Data Min Knowl Disc 1998;2(2):121–67.
[22] Zhang SC, Li XL, Zong M, Zhu XF, Wang RL. Efficient kNN Classification With Different Numbers of Nearest Neighbors. Ieee T Neur Net Lear 2018;29(5):1774–85.
[23] Safavian SR, Landgrebe D. A Survey of Decision Tree Classifier Methodology. Ieee T Syst Man Cyb 1991;21(3):660–74.
[24] Zolnierek A, Rubacha B. The empirical study of the naive Bayes classifier in the case of Markov chain recognition task. Computer Recognition Systems, Proceedings 2005:329–36.
[25] Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens 2005;26(1):217–22.
[26] Pregibon D. Logistic-Regression Diagnostics. Ann Stat 1981;9(4):705–24.
[27] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Adv Neur In 2017;30.
[28] Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. PLoS ONE 2021;16(2):e0246306.
[29] Li DH, Wald R, Blum D, et al. Predicting mortality among critically ill patients with acute kidney injury treated with renal replacement therapy: Development and validation of new prediction models. J Crit Care 2020;56:113–9.
[30] Ponce D, de Andrade LGM, Granado RC, Ferreiro-Fuentes A, Lombardi R. Development of a prediction score for in-hospital mortality in COVID-19 patients with acute kidney injury: a machine learning approach. Sci Rep 2021;11(1):24439.

Fig. 5. SHAP force analysis and Local Interpretable Model-Agnostic Explanations (LIME) algorithm for explaining individual's prediction results. Screenshot of the death prediction in patients with AKI. (A) and (B) present a deceased case with SHAP force analysis and the LIME algorithm, respectively. (C) and (D) present a deceased case with SHAP force analysis and the LIME algorithm, respectively. (A) and (C), the bars in red and blue represent risk factors and protective factors, respectively; longer bars indicate greater feature importance. (B) and (D), the left part of the figure shows predicted results using LIME. The middle part presents the top 8 variables that had the greatest impact on survival or death from top to bottom. The length of the bar for each feature indicates the importance (weight) of that feature in making the prediction. A longer bar indicates a feature that contributes more to survival or death. The right panel shows the critical values of these 8 variables when they had the greatest impact on survival or death. GCS, Glasgow Coma Scale; BUN, blood urea nitrogen; RR, respiratory rate; CVD, cerebrovascular disease; MV, mechanical ventilation; SBP, systolic blood pressure. SHAP, SHapley Additive explanation; LIME, Local Interpretable Model-Agnostic Explanations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

[31] Ohnuma T, Uchino S. Prediction Models and Their External Validation Studies for Mortality of Patients with Acute Kidney Injury: A Systematic Review. PLoS ONE 2017;12(1):e0169341.
[32] Luo M, Yang Y, Xu J, et al. A new scoring model for the prediction of mortality in patients with acute kidney injury. Sci Rep 2017;7(1):7862.
[33] Skarupskiene I, Adukauskiene D, Kuzminskiene J, et al. Mortality prediction in patients with acute kidney injury requiring renal replacement therapy after cardiac surgery. Medicina (Kaunas) 2017;53(4):217–23.
[34] Mo M, Pan L, Huang Z, Liang Y, Liao Y, Xia N. Development and Validation of a Prediction Model for Survival in Diabetic Patients With Acute Kidney Injury. Front Endocrinol (Lausanne) 2021;12:737996.
[35] Shkirkova K, Saver JL, Starkman S, et al. Frequency, Predictors, and Outcomes of Prehospital and Early Postarrival Neurological Deterioration in Acute Stroke: Exploratory Analysis of the FAST-MAG Randomized Clinical Trial. JAMA Neurol 2018;75(11):1364–74.
[36] Abdallah A, Demaerschalk BM, Kimweri D, et al. A comparison of the Full Outline of Unresponsiveness (FOUR) and Glasgow Coma Scale (GCS) Scores in Predicting Mortality Among Patients with Reduced Level of Consciousness in Uganda. Neurocrit Care 2020;32(3):734–41.
[37] Kellum JA, Sileanu FE, Murugan R, Lucko N, Shaw AD, Clermont G. Classifying AKI by Urine Output versus Serum Creatinine Level. J Am Soc Nephrol 2015;26(9):2231–8.