Scholar Publishing Group
International Journal of World Medicine
https://doi.org/10.38007/IJWM. 2026.070107.
ISSN 2789-7133 Vol. 7, Issue 1: 66-81

Nursing-Oriented Risk Prediction and Response to Sepsis-Associated Acute Kidney Injury Using Interpretable Machine Learning and Delphi Consensus

Yang Ding[1], Minerva Bravo De Ala[1,*] and Ying Zhang[2]

[1]*School of Nursing, Philippine Women's University, Manila 1004, Philippines*

[2]*School of Emergency and Trauma, Hainan Medical University, Haikou 571199, Hainan, China*

[*]*Corresponding author*

Keywords: Sepsis-associated acute kidney injury; Machine learning; Risk prediction; Delphi technique; Critical care nursing

Abstract: In fact, SA-AKI is usually detected only when there are signs of renal impairment; hence, little room remains for nurses to carry out prophylactic evaluation and intervention measures in advance. Therefore, this investigation established a readily understandable, nurse-centered model to predict early SA-AKI risks. External validation was then performed. Finally, based on the model's predictions, a ladder-type nursing response mode was created accordingly. Then they used routinely recorded measurements taken during their patients' first 24 hours in the ICU to try to predict which of these individuals developed new cases of SA-AKI sometime during their first 48 hours after entering the ICU. They developed models using the eICU Collaborative Research Database (eICU) and performed internal validation on this database alone. Then they externally validated their models using two other datasets: the Medical Information Mart for Intensive Care IV (MIMIC-IV) database and a cohort from the First Affiliated Hospital of Hainan Medical University (FAH-HMU). There were four kinds of feature-selection approaches involved, taken together. Nine types of machine-learning algorithm were fine-tuned by means of Bayesian optimization and then compared. To explain what can be predicted by the last model chosen, the method adopted was Shapley additive explanations. Regarding the derivation of possible indicators from among candidates found while developing the model itself, a two-stage Delphi consultation followed about their clinical value, whether nurses could use them, and whether they fit easily with existing ways of working. It involved 15,767 patients from the eICU database, 12,842 subjects from MIMIC-IV, and 932 people from the FAH-HMU group. Here too, the best results were found for the Extreme Gradient Boosting algorithm: the area under the receiver operating characteristic curve was 0.798 (internal validation), 0.738 (MIMIC-IV), and 0.727 (institutional sample). There was also good calibration, and decision-curve analysis provided additional reasons to advocate using this method for early risk stratification. All fifteen experts answered both Delphi questionnaires, and the authority coefficient value was 0.87. Delphi consensus retained nine objective indicators and added capillary refill time as an additional nurse's bedside observation item. On this basis, three kinds of pathways according to high-, moderate-, and

low-risk levels were set forth, along with specific requirements for monitoring, reevaluation, and escalation, respectively. It produces a nursing response scheme consisting of multicohort-validated and explainable risk prediction combined with expert consensus on the levels of nursing activity required. Thus, a certain structure exists regarding what needs to happen in terms of early detection of SA-AKI risk, bedside monitoring, and risk-based nursing care in intensive-care practice.

1. Introduction

SA-AKI occurs quite often in people who are critically ill, but unfortunately its outcome is poor. About 21.6% of hospitalized adults develop AKI during their hospital stay, while this proportion rises to 40%–57.3% among ICU patients [1,2]. When AKI occurs together with sepsis, more renal replacement therapies are needed, longer ICU stays and higher costs are observed, and mortality also increases; moreover, even if patients survive, they remain at risk of developing CKD and cardiovascular diseases later on [3,4]. Therefore, early detection of SA-AKI is important because the kidneys may already have been injured even though diagnostic criteria have not yet been met.

It depends on what KDIGO means by serum creatinine and urine output [5]. Serum creatinine may take time to increase and also varies according to muscle bulk, hydration, and sepsis-induced shifts in metabolism. In theory, one could use urine output instead, but here again there are problems. For example, fluid balance matters, as do diuretics, hemodynamics, and many other causes unrelated to kidney function altogether. Therefore, ICU nurses are aware of creatinine values and amounts of urine passed, along with vital signs, perfusion, oxygenation, fluid balance, treatments received, and responses noted. They have no difficulty obtaining this type of information; the real issue arises from having too much of it in too little space and needing to determine exactly what everything taken collectively actually signifies at the present moment.

There is much more freedom in how machine-learning methods represent relationships among EHR features compared to what is possible using fixed scoring algorithms such as APACHE II, SOFA, and SAPS II [6–8]. Some success has been reported in predicting AKI by gradient-boosting machines, and SHAP may provide useful information about which features drove particular predictions [9–13]. But discrimination alone does not suffice if there is to be clinical value; external validity, calibration, decision-analytic benefit, intelligibility, and ease of integration into nursing practice all need attention as well [14,15].

And here we coupled model development with clinical translation. That is, models were trained on one multicentre ICU database and tested on two others; then came SHAP analysis, followed by two rounds of Delphi consultation where experts rated items according to certain criteria (clinical relevance, nursing feasibility, workflow fit). In this way they hoped to reach three goals: first, to determine a patient's chances of developing SA-AKI later ("before" the outcome window); second, to identify the measurements responsible for producing this degree of risk; third, to suggest what kind of nursing response might suit each situation ("proportionate" responses). Finally, they wished to present their result as an aid to decision-making about monitoring patients and escalating care, not as a machine that makes diagnoses by itself.

2. Materials and Methods

2.1 Design, Data Sources, and Participants

A multiphase mixed-methods design was used. Phase one comprised retrospective model development, internal validation, external validation, and interpretation. Phase two used Delphi consensus to refine the model-derived features and formulate a nursing response pathway.

The eICU Collaborative Research Database (eICU) was used for development and internal validation [16]. External validation used the Medical Information Mart for Intensive Care IV (MIMIC-IV), version 3.0 [17], and the First Affiliated Hospital of Hainan Medical University (FAH-HMU) intensive care unit (ICU) cohort (2021–2024). Eligible patients were adults in their first ICU admission who met Sepsis-3 criteria within 24 hours, stayed at least 48 hours, and had sufficient data to define the outcome and predictors [18]. We excluded pre-existing end-stage renal disease, chronic renal replacement therapy, kidney transplantation, and acute kidney injury present on admission or during the first 24 hours.

2.2 Outcome and Predictors

The primary outcome was new SA-AKI of any KDIGO stage during hours 24–48 after ICU admission [5]. It was present when creatinine increased by at least 0.3 mg/dL within 48 hours, reached 1.5 times baseline, or urine output remained below 0.5 mL/kg/h for at least six consecutive hours. Baseline creatinine was the lowest value in the preceding seven days when available; otherwise, the admission value was used.

Candidate predictors were measured during the first 24 hours and included demographics, vital-sign summaries, blood gas and laboratory results, comorbidities, SOFA and other severity measures, mechanical ventilation, vasopressor and diuretic exposure, fluid input, and urine output. Elixhauser comorbidities were derived from diagnostic codes [19]. Variables were mapped across databases and converted to common units. Records or variables beyond prespecified missingness limits were removed; remaining values were imputed by chained equations using random-forest models [20,21].

2.3 Feature Selection, Modeling, and Interpretation

Four procedures were carried out on the whole eICU training set: LASSO stability selection [22], Boruta [23], recursive feature elimination with cross-validation (RFE-CV) [24], and stepwise regression based on the Akaike information criterion [22–24]. A variable then joined the final group if three or more methods selected it among their outputs. By using votes among methods instead of relying on a single method, some flexibility was built into the scheme.

There were nine algorithms altogether: logistic regression, random forest, support vector machine (SVM), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), multilayer perceptron (MLP), AdaBoost, CatBoost, and explainable boosting machine (EBM) [9,25–27]; there was also Bayesian optimization with five-fold cross-validation restricted to the training portion of eICU [28]. Selected settings were then tested on the remainder of eICU and, without retraining, also on MIMIC-IV and FAH-HMU.

There was evaluation concerning discrimination, calibration, and clinical usefulness. For this reason, we computed the values of AUC, Brier score, accuracy, sensitivity, specificity, PPV, NPV, and F1-score. Calibration curves showed relations between predicted and actual probabilities; decision curve analysis yielded information about net benefit at various threshold probabilities [29]. SHAP revealed which variables contributed most overall, what their signs were, and also allowed us to examine specific cases. In addition, we determined the top predictors according to both XGBoost and LightGBM [12,13]. Finally, partial dependence plots helped to detect any possible nonlinearities.

2.4 Delphi Consultation and Statistics

Then came two rounds of electronic Delphi procedures according to model interpretation [30]. Purposive sampling identified ICU nurses, nursing educators, nursing researchers, critical-care physicians, or emergency physicians who had more than five years' relevant experience and held either an associate senior title or a master's degree. They rated clinical importance, nursing feasibility, and workflow applicability on five-point Likert scales and also wrote comments. A subsequent round allowed participants to learn what others thought about these issues and to respond again after seeing the changes made to features and responses based on their own comments. Because the study concerned clinical decision support [31], consultation followed published advice concerning the use of the method [32].

Preliminary consensus required a mean importance score of at least 4.0 and a coefficient of variation below 0.25. Round-two retention required a mean of at least 4.0 and a coefficient of variation below 0.20. Expert authority was the mean of judgment and familiarity coefficients; Kendall's W assessed concordance. Continuous variables were summarized as mean (standard deviation) or median (interquartile range), and categorical variables as n (%). Between-group tests were selected according to distribution and cell size. Statistical tests were two-sided. Generative AI tools were not used for data extraction, feature selection, model training, statistical analysis, or Delphi procedures; their use in manuscript preparation is disclosed in the AI Use Statement.

2.5 Ethics

The public databases contain de-identified records and were accessed after required training and data-use agreements [16,17]. The Ethics Committee of the First Affiliated Hospital of Hainan Medical University approved the FAH-HMU analysis (2026-KYL-021) and waived individual consent for the retrospective, de-identified cohort. Delphi participation was voluntary and responses were summarized at group level.

3. Results

3.1 Cohorts and Baseline Characteristics

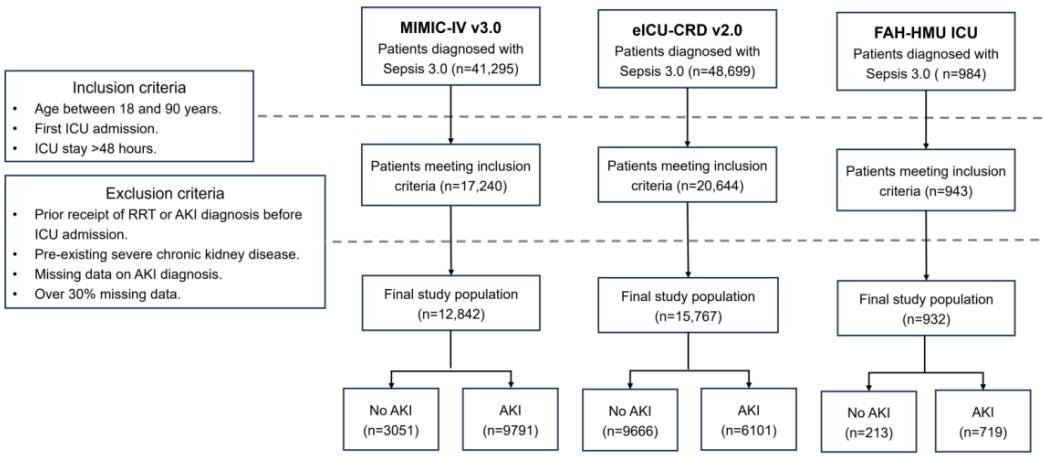


Figure 1. Patient selection in the eICU, MIMIC-IV, and FAH-HMU ICU cohorts. Abbreviations: AKI, acute kidney injury; eICU, eICU Collaborative Research Database; FAH-HMU, First Affiliated Hospital of Hainan Medical University; ICU, intensive care unit; MIMIC-IV, Medical Information Mart for Intensive Care IV.

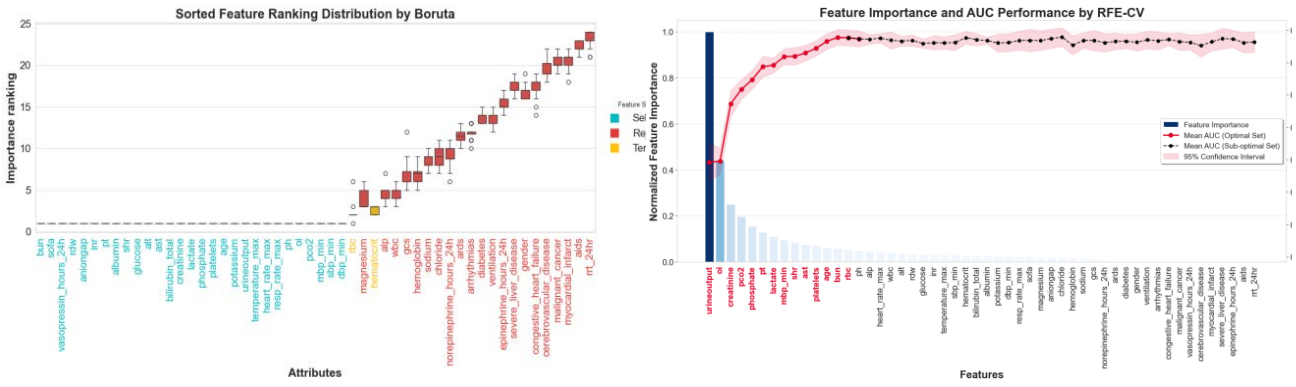
Uniform eligibility rules reduced 48,699 eICU Sepsis-3 admissions to 15,767 patients; 6,101 developed SA-AKI and 9,666 did not. MIMIC-IV contributed 12,842 of 41,295 screened admissions (9,791 SA-AKI; 3,051 no SA-AKI). The FAH-HMU cohort contributed 932 of 984 admissions (719 SA-AKI; 213 no SA-AKI) (Figure 1). In eICU, patients who developed SA-AKI were older and had lower urine output and blood pressure, higher lactate, creatinine, blood urea nitrogen, and SOFA scores, and more frequent mechanical ventilation (Table 1).

Table 1. Selected baseline characteristics of the eICU cohort.

Characteristic	No SA-AKI (n = 9,666)	SA-AKI (n = 6,101)	*p* value
Age (years)	62.5 (16.3)	64.6 (15.2)	<0.001
Urine Output (mL)	2040 (1330)	1220 (1210)	<0.001
Ventilation	2946 (30.5%)	2145 (35.2%)	<0.001
Maximum Temperature (°C)	38.2 (0.51)	39.4 (0.65)	0.00183
Minimum Systolic Blood Pressure (mmHg)	88.1 (19.7)	83.2 (19.4)	<0.001
Minimum Mean Blood Pressure (mmHg)	60.1 (13.6)	56.3 (13.4)	<0.001
Oxygenation Index	1750 (4990)	1650 (4920)	<0.001
Phosphate (mg/dL)	3.43 (1.10)	3.85 (1.44)	<0.001
Lactate (mmol/L)	2.62 (2.05)	3.30 (2.83)	<0.001
Blood Urea Nitrogen (mg/dL)	29.6 (24.0)	34.3 (24.8)	<0.001
Creatinine (mg/dL)	1.41 (1.27)	1.86 (1.63)	<0.001
Total Bilirubin (mg/dL)	1.09 (1.76)	1.50 (2.77)	<0.001
Aspartate Aminotransferase (U/L)	111 (403)	229 (1000)	<0.001
SOFA Score	5.93 (2.83)	7.06 (3.07)	<0.001

Note. Values are mean (SD) or n (%). The table reports prespecified clinically relevant characteristics; the model used the complete harmonized candidate set. SA-AKI, sepsis-associated acute kidney injury; SD, standard deviation; SOFA, Sequential Organ Failure Assessment.

3.2 Feature Selection



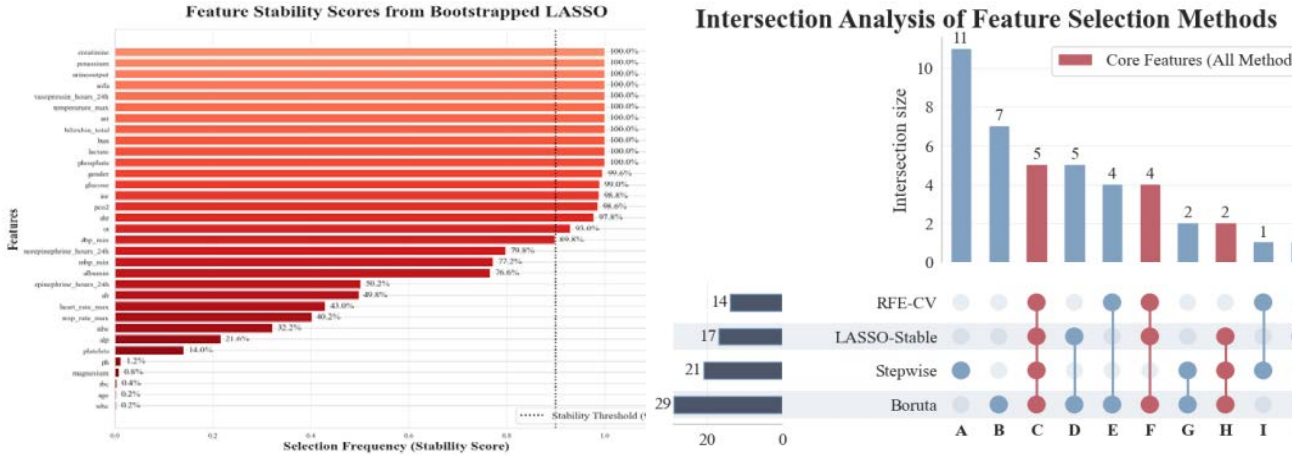


Figure 2. Feature-selection diagnostics: (a) Boruta importance ranking; (b) recursive feature elimination with cross-validation (RFE-CV); (c) least absolute shrinkage and selection operator (LASSO) stability selection; and (d) overlap across the four selection procedures.

Table 2. Stepwise regression feature-selection results across forward, backward, and bidirectional procedures.

Feature	Forward	Backward	Bidirectional	Vote count
Anion gap	TRUE	TRUE	TRUE	3
Arrhythmias	TRUE	TRUE	TRUE	3
Total bilirubin	TRUE	TRUE	TRUE	3
BUN	TRUE	TRUE	TRUE	3
Cerebrovascular disease	TRUE	TRUE	TRUE	3
Congestive heart failure	TRUE	TRUE	TRUE	3
Creatinine	TRUE	TRUE	TRUE	3
Diabetes	TRUE	TRUE	TRUE	3
Nephrotoxic score	TRUE	TRUE	TRUE	3
Malignant cancer	TRUE	TRUE	TRUE	3
Mechanical ventilation	TRUE	TRUE	TRUE	3
Urine output	TRUE	TRUE	TRUE	3
OI	TRUE	TRUE	TRUE	3
Phosphate	TRUE	TRUE	TRUE	3
RBC	TRUE	TRUE	TRUE	3
PTT	TRUE	TRUE	TRUE	3
RRT within 24 h	TRUE	TRUE	TRUE	3
RDW	TRUE	TRUE	TRUE	3
Severe liver disease	TRUE	TRUE	TRUE	3
Maximum temperature	TRUE	TRUE	TRUE	3
Lactate	FALSE	FALSE	TRUE	1
Epinephrine exposure within 24 h	FALSE	FALSE	TRUE	1
Platelets	FALSE	FALSE	TRUE	1
Minimum MBP	FALSE	FALSE	TRUE	1
Minimum SBP	FALSE	FALSE	TRUE	1
Sodium	FALSE	FALSE	TRUE	1

Note. TRUE indicates retention and FALSE indicates exclusion. ALP, alkaline phosphatase; BUN, blood urea nitrogen; MBP, mean blood pressure; OI, oxygenation index; PTT, partial

thromboplastin time; RBC, red blood cell count; RDW, red cell distribution width; RRT, renal replacement therapy; SBP, systolic blood pressure.

Boruta retained 29 variables, recursive feature elimination identified a 14-variable optimum, stability selection retained 17 variables at a 90% probability cutoff, and the stepwise procedures converged on 21 variables (Table 2). The four-method vote yielded 11 core features: blood urea nitrogen (BUN), creatinine, oxygenation index (OI), phosphate, urine output, aspartate aminotransferase (AST), lactate, partial pressure of carbon dioxide, stress hyperglycemia ratio (SHR), total bilirubin, and maximum temperature (Figure 2).

3.3 Model Performance, Calibration, and Utility

Random forest attained the largest training AUC (0.934), yet it lost discrimination when validated and thus appears overfitted. Some algorithms cluster together in obtaining nearly identical AUCs near 0.80 for the eICU internal sample. XGBoost obtains an AUC of 0.798, a Brier score of 0.182, accuracy of 0.734, sensitivity of 0.716, specificity of 0.746, positive predictive value of 0.640, negative predictive value of 0.806, and an F1 score of 0.676. Therefore, XGBoost was chosen for implementation because, although none of these metrics is best on every dataset or measure, overall there seems to be a good balance among discrimination, sensitivity, calibration, net benefit, and interpretability.

There was also some external AUC of XGBoost: 0.738 in MIMIC-IV and 0.727 in FAH-HMU (Table 3). Although logistic regression took first place in two kinds of external cohorts, XGBoost still achieved medium discrimination and produced intelligible reasons behind every prediction. Decision curves looked good from about 20% to nearly 80% on the probability axis in each validation sample. Internally, calibration was fairly near the diagonal, while the Brier value was 0.296; therefore, there might have been calibration problems caused by differences among populations when transferred to MIMIC-IV (Figures 3 and 4).

Table 3. Then there is discrimination and calibration across validation cohorts.

Model	eICU AUC	eICU Brier	MIMIC-IV AUC	MIMIC-IV Brier	FAH-HMU AUC	FAH-HMU Brier
Logistic Regression	0.7722	0.1989	0.7585	0.2639	0.7529	0.2694
Random Forest	0.7984	0.1783	0.7327	0.3057	0.7197	0.3234
SVM	0.7899	0.1783	0.7412	0.3488	0.7046	0.3782
XGBoost	0.7981	0.1817	0.7375	0.2958	0.7266	0.3137
LightGBM	0.7962	0.1815	0.7336	0.3013	0.7150	0.3177
MLP	0.7935	0.1773	0.7529	0.3616	0.7348	0.3633
AdaBoost	0.7903	0.2346	0.7472	0.2534	0.7257	0.2559
CatBoost	0.7975	0.1816	0.7370	0.3068	0.7128	0.3205
EBM	0.7985	0.1743	0.7453	0.3551	0.7282	0.3725

Note. AUROC, short for AUC (area under the receiver operating characteristic curve); the abbreviation “eICU” refers to “the eICU Collaborative Research Database,” while “FAH-HMU” indicates “the First Affiliated Hospital of Hainan Medical University.” “MIMIC-IV” is an abbreviation for “Medical Information Mart for Intensive Care IV.” EBM stands for explainable boosting machine, MLP for multilayer perceptron, and SVM for support vector machine.

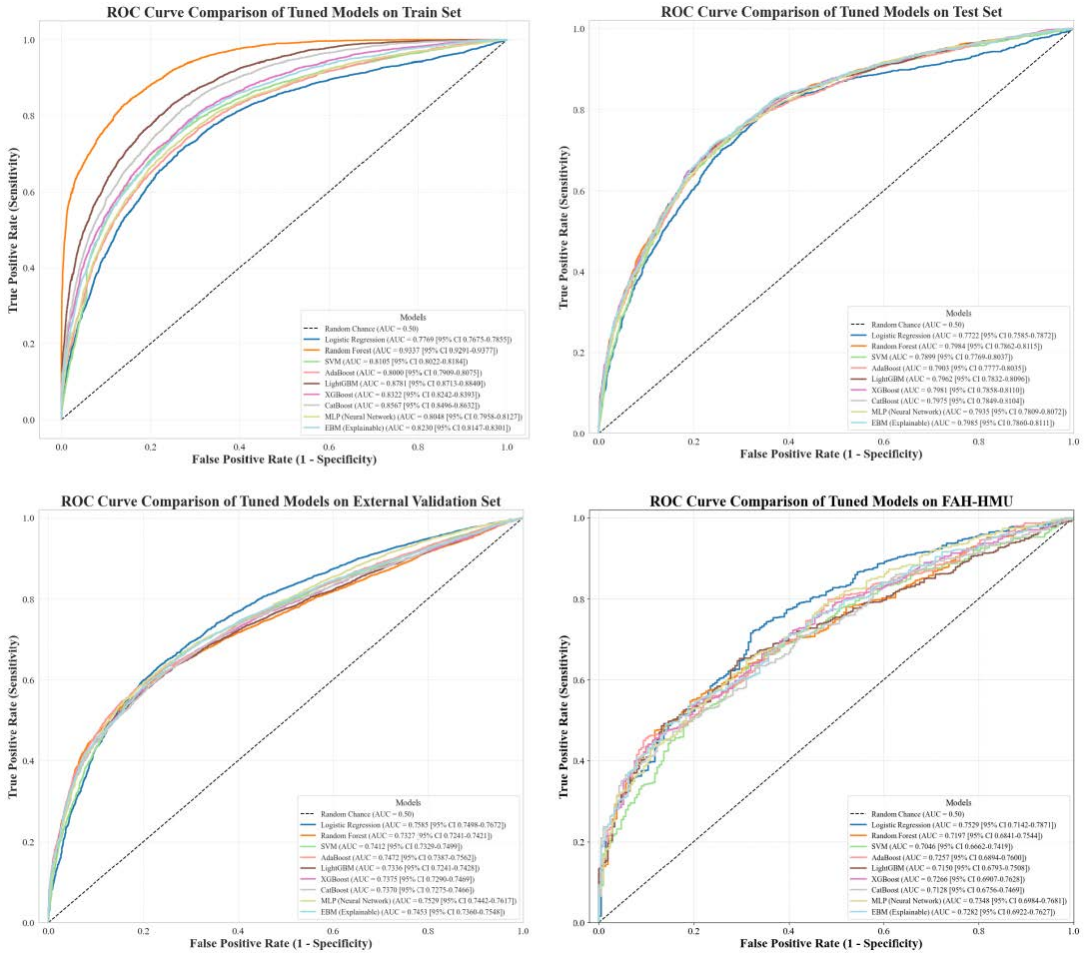


Figure 3. Receiver operating characteristic (ROC) curves for the nine candidate models.

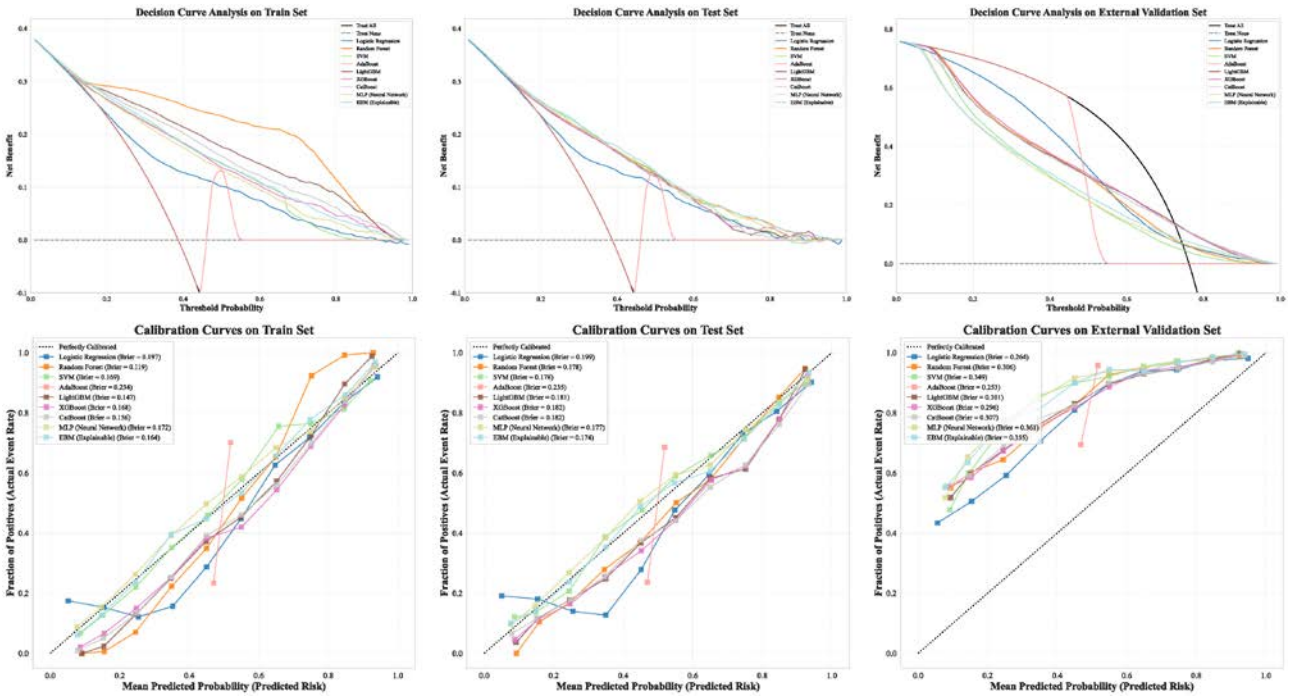


Figure 4. Decision Curves and Calibration Plots across Development and Validation Cohorts

3.4 Model Interpretation

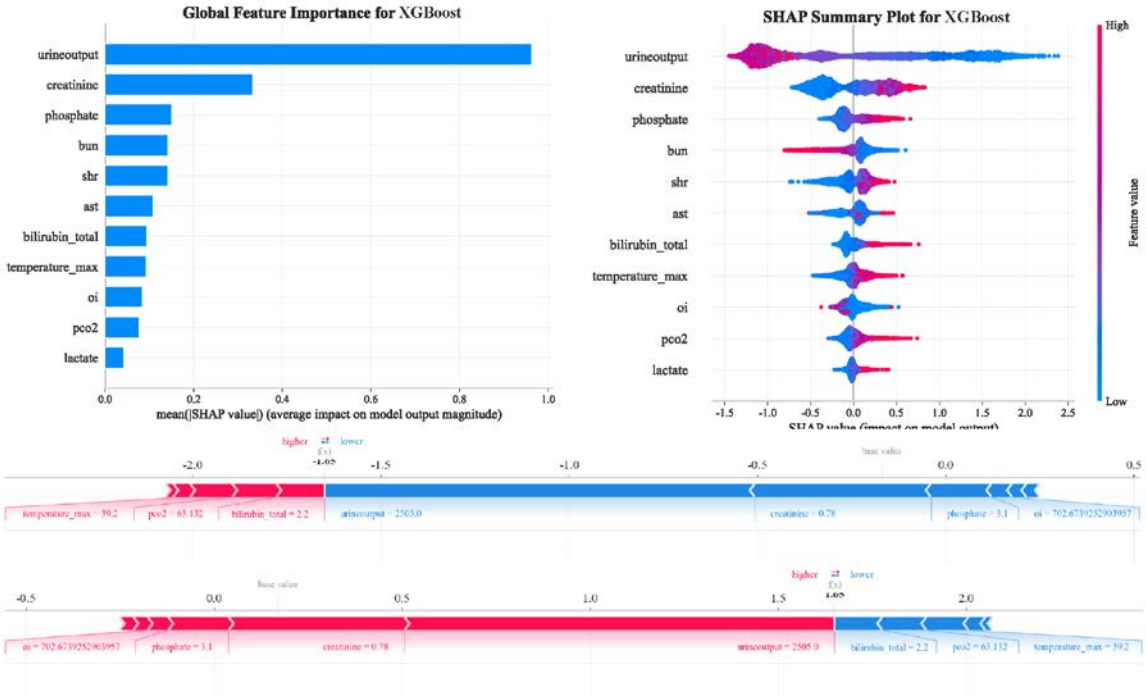


Figure 5. Then comes global and patient-level Shapley additive explanations (SHAP) interpretation of the Extreme Gradient Boosting (XGBoost) model. Abbreviations: AST, aspartate aminotransferase; BUN, blood urea nitrogen; OI, oxygenation index; PCO_2, partial pressure of carbon dioxide; SHR, stress hyperglycemia ratio.

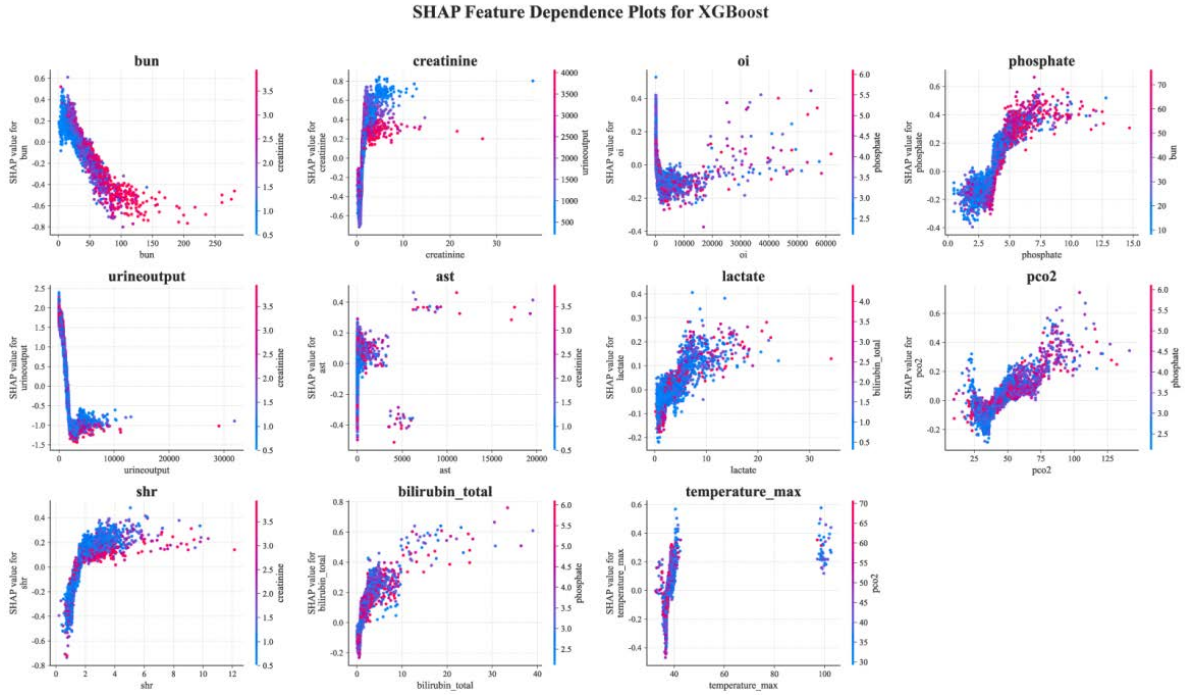


Figure 6. Partial Dependence Plots for Some Important Continuous Predictors of SA-AKI Risk. Abbreviations: AST, aspartate aminotransferase; BUN, blood urea nitrogen; OI, oxygenation index; PCO_2, partial pressure of carbon dioxide; SHR, stress hyperglycemia ratio.

In fact, urine output had the highest mean absolute SHAP value among all SHAP values. Creatinine, phosphate, blood urea nitrogen, stress hyperglycemia ratio, aspartate aminotransferase, total bilirubin, lactate, oxygenation index, partial pressure of carbon dioxide, and maximum temperature were contributors. Smaller urine outputs and larger values of creatinine, phosphate, blood urea nitrogen, and stress hyperglycemia ratio usually shifted prediction results toward SA-AKI. Individual plots showed the specific outcomes for each patient (Figure 5).

There was some evidence of partial dependence on estimated risk from below roughly 1,500 mL down to zero daily urine volume, whereupon there was probably more of a jump. There was also likely greater risk above approximately 1.2 mg/dL serum creatinine and increasing levels of phosphate and BUN (Figure 6). XGBoost and LightGBM had three out of five features in common; namely, these were once again the top three predictors: urine output, creatinine, and lactate. Therefore, although there is still room for considerable algorithmic difference, some recurrently important clinical variables again seem to emerge.

3.5 Delphi Consensus and Nursing Response

In fact, all fifteen experts responded in both rounds (response rate = 100%). Of these, ICU clinical nurses made up 66.7%, while 80.0% held an associate senior title or above; additionally, 80.0% had obtained a master's degree or above, and 80.0% had more than ten years of work experience. Their judgment coefficient was 0.88, and their familiarity coefficient was 0.86; therefore, their authority coefficient was 0.87.

Phase 1 supported creatinine, BUN, lactate, urine output, oxygenation index, $PaCO_2$, total bilirubin, and some selected inflammatory and metabolic measures. With further feedback, the final package consisted of nine objective indicators plus CRT. Experts introduced CRT on their own initiative as "one more thing" (something quickly done at the bedside) concerning perfusion (see Table 4). They dropped phosphate because they could not obtain enough samples; they removed SHR because they found this indicator hard to explain; finally, they excluded AG since someone wanted to act upon it immediately. There was substantial agreement regarding feature weightings ($W = 0.218$, $\chi^2 = 29.43$, $p < 0.001$); there was also considerable consensus about fitting into practice routines ($W = 0.241$, $\chi^2 = 25.31$, $p < 0.001$); moreover, they reached sufficient agreement concerning the reaction system itself ($W = 0.285$, $\chi^2 = 34.20$, $p < 0.001$).

Table 4. Delphi final consensus for the nursing early warning feature set.

Category	Indicator	Type	Round 2 mean ± SD	CV
Renal function and elimination	Creatinine	Objective clinical indicator	4.98 ± 0.14	0.03
Renal function and elimination	Urine output	Objective clinical indicator	4.93 ± 0.26	0.05
Renal function and elimination	Blood urea nitrogen	Objective clinical indicator	4.87 ± 0.35	0.07
Perfusion and metabolism	Lactate	Objective clinical indicator	4.93 ± 0.26	0.05
Perfusion and metabolism	Capillary refill time	Nursing observation indicator	4.80 ± 0.41	0.09
Respiratory and acid-base status	Oxygenation index	Objective clinical indicator	4.40 ± 0.63	0.14

Category	Indicator	Type	Round 2 mean ± SD	CV
Respiratory and acid-base status	Partial pressure of carbon dioxide	Objective clinical indicator	4.47 ± 0.51	0.11
Inflammation and liver function	Maximum body temperature	Objective clinical indicator	4.13 ± 0.52	0.12
Inflammation and liver function	Aspartate aminotransferase	Objective clinical indicator	4.07 ± 0.59	0.14
Inflammation and liver function	Total bilirubin	Objective clinical indicator	4.00 ± 0.65	0.16

Note. CV, coefficient of variation; SD, standard deviation. 'Objective' refers to an electronically recorded clinical measure; capillary refill time is a bedside finding made by nurses.

It advised assessing once per nursing shift and having a fixed display location within the nursing electronic record. This display will show what tier someone is in, the variables that contributed most to making this calculation, their recent trends, and reference ranges (Figure 7). Being high-risk means there is a need for close monitoring, checking whether urine output was confirmed, reviewing inputs and outputs and perfusion parameters more frequently, and acting early if there appear to be problems that require a doctor's attention, etc.; being moderate-risk demands prioritized trend review and intensified observation; low-risk persons continue routine monitoring with reassessment for movement between tiers (Figure 7).

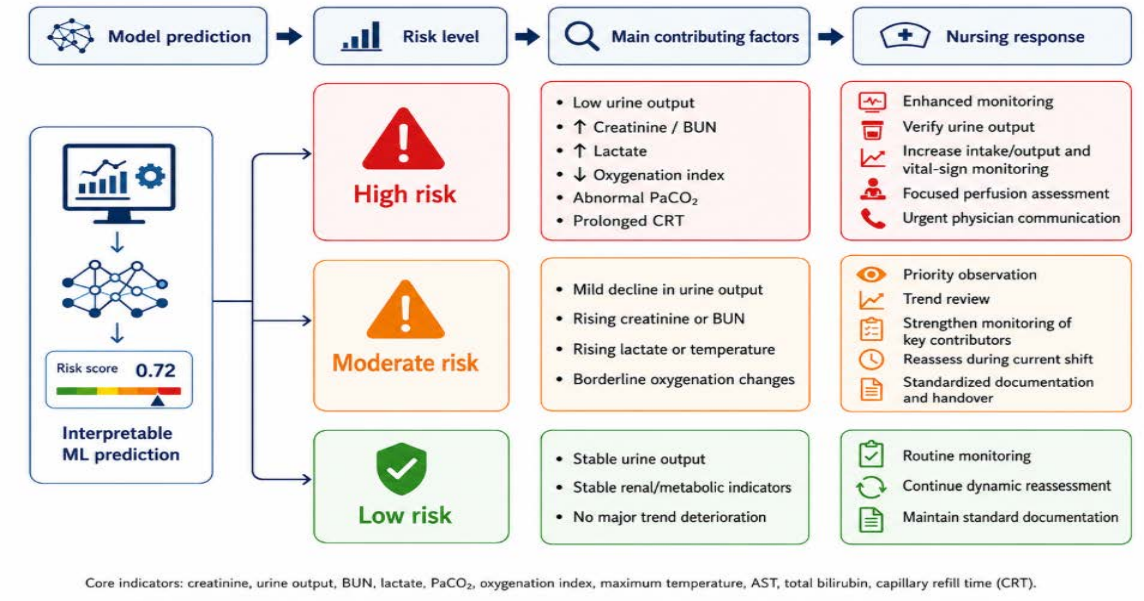


Figure 7. Shifting displays and stepwise nursing responses according to SA-AKI risk stratification
Abbreviations: AST indicates aspartate aminotransferase; BUN indicates blood urea nitrogen; CRT indicates capillary refill time; ML indicates machine learning; OI indicates oxygenation index; $PaCO_2$ indicates partial pressure of arterial carbon dioxide

4. Discussion

4.1 Main Findings

It was thus possible to link together three kinds of work that are frequently reported separately: prediction, explanation, and clinical translation. XGBoost gave moderately good and fairly constant

discrimination throughout one internal sample and two external samples, although logistic regression had the largest external AUCs. Choices were made on the basis of the entire performance profile and also on grounds of explaining individual predictions, not merely by asserting that one model worked best overall. The proposed use is early risk stratification before KDIGO criteria are fully expressed.

It is thus presented as SA-AKI multisystem dysfunction. Urine output, creatinine, and blood urea nitrogen refer to renal clearance; lactate and capillary refill time concern perfusion; oxygenation index and carbon dioxide relate to respiration and acid-base balance; temperature, aspartate aminotransferase, and bilirubin point to signs of inflammation and liver involvement. Some changes occurred between the algorithmic list derived by design and what finally emerged from Delphi deliberations. In one case, phosphate proved to be a good predictor but then had to be removed because some members questioned whether they could measure it often enough or knew what nurses should actually do if values were abnormal. For another indicator, capillary refill time, precisely the reverse happened. Although absent among the items recorded in a structured fashion for model-building purposes, it was reinstated for exactly this reason: nurses can obtain these numbers quickly at the patient's bedside.

4.2 Clinical Interpretation and Nursing Use

That is why attention needs to be paid to the nonlinearity of the urine-output relation. In fact, this applies generally: there is no reason to treat totals in isolation as if they were set once and for all by some fixed administration policy. They need to be followed along with fluid inputs, modifications of vasopressors, measures relating to perfusion and metabolism, and everything else. As mentioned above, creatinine entered into the diagnosis, but inside the model it participated in making up a changing picture instead of serving alone beyond a certain level [5]. Obviously, care must be taken in practice not to produce too many messages, since these can lead to warning apathy [33]. Both lactic acid levels and capillary-refill times provide alternative ways of examining systemic and peripheral perfusion. However, they cannot capture the whole picture of hemodynamics. That was demonstrated when treating septic-shock patients according to guidelines [34].

Explanation might also make escalation itself more worthwhile. Here is one possible picture. There is some estimated risk, along with falling urine output, rising creatinine, and increased lactate. Thus there is something definite to raise. Then comes consideration of hemodynamics, nephrotoxic exposure, whether tests need repeating, and what renal support options exist, depending on exactly what has happened clinically. Again, this kind of help resembles the provision of appropriate information “just in time” mentioned above as part of CDS [31].

Panel preference was for shift-based display over repeated real-time notifications, on the grounds that such a timetable satisfies handovers and set times for reconsideration without creating yet another cause of alarm fatigue [33], and response levels prescribe some minimal activity without telling people what they have to do; novices get specific instructions to follow, and more knowledgeable nurses modify this process according to their judgment about patients’ conditions generally. The relationship between algorithm and health-care provider is supposed to produce improvement rather than replacement [35].

4.3 Strengths and Limitations

Advantages consist of multicenter development and external testing in an international database as well as in a Chinese institutional cohort; examination of nine algorithms; attention to issues beyond AUC; SHAP interpretation at cohort and individual levels; Delphi consultation about

whether nurses can actually implement what the models propose; and continuation of the study beyond mere prediction to consideration of exactly what happens next in practice.

Several problems still need attention. First, retrospective structured records carry their own problems of measurement error, coding differences, residual confounding, and missing nursing context. Second, there were no free-text notes or serial bedside observations available. Third, performance varied among cohorts, as did calibration; thus, local recalibration may be necessary. Fourth, the Delphi panel was small and mainly drawn from Chinese institutions; its recommendations may not transfer unchanged to other staffing models or electronic records. Fifth, the response framework has not been prospectively tested for recognition time, SA-AKI incidence, workload, alarm fatigue, renal replacement therapy, or patient outcomes. Sixth, model drift is also possible as sepsis care and documentation change. Therefore, before any decision about routine adoption, more research will probably first need to occur. In particular, further trials should initially take place in several sites simultaneously, along with tests of ease of use, performance measurements, and systematic collection of nursing observations. [36]

5. Conclusion

Thereupon, an interpretable model to predict SA-AKI probability throughout the first 24 hours of intensive care itself arose; it then underwent testing in two separate groups. XGBoost yielded satisfactory discrimination together with an intelligible risk picture; Delphi agreement turned model-based features into nine clear-cut signs plus one measurement value concerning capillary-refill time divided among three kinds of nurse reaction. In summary, therefore, there appears here something that can serve as a foundation for surveillance and stepping up treatment, although it does not amount to a diagnostic guideline by itself, nor does anything yet indicate whether this scheme will change what nurses end up doing or benefit patients.

Acknowledgement

We thank the School of Nursing at Philippine Women’s University for academic guidance; Hainan Medical University and the First Affiliated Hospital of Hainan Medical University for clinical data management and local validation support; the eICU and MIMIC-IV teams for maintaining the public databases; and the Delphi panelists for their clinical judgments.

Funding Statement

The author(s) received no specific funding for this study.

Availability of Data and Materials

Credentialed researchers may access eICU and MIMIC-IV through PhysioNet after completing the required training and data-use agreements. The FAH-HMU cohort and Delphi records cannot be released publicly because of ethics, privacy, and institutional governance requirements. Reasonable requests may be submitted to the corresponding author and require institutional approval.

Ethics Approval

The Ethics Committee of the First Affiliated Hospital of Hainan Medical University approved the FAH-HMU cohort (2026-KYL-021) in accordance with applicable institutional and ethical standards. Informed consent was waived for the retrospective, de-identified cohort. Public-database

access complied with the applicable training and data-use agreements. Delphi participation was voluntary.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

AI Use Statement

ChatGPT (OpenAI) was used solely for English-language editing and abbreviation checking. All output was reviewed and verified by the authors, who assume full responsibility for the accuracy, originality, and integrity of the manuscript.

References

[1] *Hoste, E. A. J., Bagshaw, S. M., Bellomo, R., Cely, C. M., Colman, R., Cruz, D. N., Edipidis, K., Forni, L. G., Gomersall, C. D., Govil, D., Honoré, P. M., Joannes-Boyau, O., Joannidis, M., Korhonen, A.-M., Lavrentieva, A., Mehta, R. L., Palevsky, P., Roessler, E., Ronco, C., … Kellum, J. A. (2015). Epidemiology of acute kidney injury in critically ill patients: The multinational AKI-EPI study. Intensive Care Medicine, 41(8), 1411–1423. https://doi.org/10.1007/s00134-015-3934-7*

[2] *Susantitaphong, P., Cruz, D. N., Cerda, J., Abulfaraj, M., Alqahtani, F., Koulouridis, I., & Jaber, B. L. (2013). World incidence of AKI: A meta-analysis. Clinical Journal of the American Society of Nephrology, 8(9), 1482–1493. https://doi.org/10.2215/CJN.0 0710113*

[3] *Peerapornratana, S., Manrique-Caballero, C. L., Gómez, H., & Kellum, J. A. (2019). Acute kidney injury from sepsis: Current concepts, epidemiology, pathophysiology, prevention and treatment. Kidney International, 96(5), 1083–1099. https://doi.org/10.1016/j.kint.2019.05.026*

[4] *Poston, J. T., & Koyner, J. L. (2019). Sepsis associated acute kidney injury. BMJ, 364, k4891. https://doi.org/10.1136/bmj.k48 91*

[5] *Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. (2012). KDIGO clinical practice guideline for acute kidney injury. Kidney International Supplements, 2(1), 1–138. https://doi.org/10.1038/kisup.2012.1*

[6] *Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C. K., Suter, P. M., & Thijs, L. G. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine, 22(7), 707–710. https://doi.org/10.1007/BF01709751*

[7] *Knaus, W. A., Draper, E. A., Wagner, D. P., & Zimmerman, J. E. (1985). APACHE II: A severity of disease classification system. Critical Care Medicine, 13(10), 818–829. https://doi.org/10.1097/00003246-198510000-00009*

[8] *Le Gall, J. R., Lemeshow, S., & Saulnier, F. (1993). A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA, 270(24), 2957–2963. https://doi.org/10.1001/jama.1993.0351024006 9035*

[9] *Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785*

[10] Tomašev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., Mottram, A., Meyer, C., Ravuri, S., Protsyuk, I., Connell, A., Ouyang, C., Karthikesalingam, A., Nielson, C., Kohlberg, E., Mohamed, S., Clancy, B., Fontana, V., McNamee, C., … Hassabis, D. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572(7767), 116–119. https://doi.org/10.1038/s41586-019-1390-1

[11] Flechet, M., Güiza, F., Schetz, M., Wouters, P., Vanhorebeek, I., Derese, I., Gunst, J., Spriet, I., Van den Berghe, G., & Meyfroidt, G. (2017). AKI predictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: Development, validation, and comparison to serum neutrophil gelatinase-associated lipocalin. Intensive Care Medicine, 43(6), 764–773. https://doi.org/10.1007/s00134-017-4678-3

[12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30 (pp. 4765–4774).

[13] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[14] Li, J., Zhu, M., & Yan, L. (2024). Predictive models of sepsis-associated acute kidney injury based on machine learning: A scoping review. Renal Failure, 46(2), 2380748. https://doi.org/10.1080/0886022X.2024.2380748

[15] Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., Ratliff, W., & Balu, S. (2020). A path for translation of machine learning products into healthcare delivery. EMJ Innovations, 4(1), 38–45. https://doi.org/10.33590/emjinnov/19-0 0172

[16] Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data, 5, 180178. https://doi.org/10.1038/sdata.2018.178

[17] Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-W. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, 1. https://doi.org/10.1038/s41597-022-01899-x

[18] Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock. JAMA, 315(8), 801–810. https://doi.org/10.1001/jama.2016.0287

[19] Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. Medical Care, 36(1), 8–27. https://doi.org/10.1097/00005650-199801000-00004

[20] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss. v045.i03

[21] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest: Non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/bio informatics/btr597

[22] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02 080.x

[23] Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. Journal of Statistical Software, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11

[24] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389–422. https://doi.org/10.1023/A:1012487302797

[25] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 30 (pp. 3146–3154).

[26] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems, 31 (pp. 6639–6649).

[27] Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 623–631). https://doi.org/10.1145/2487575.2487579

[28] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2623–2631). https://doi.org/10.1145/3292500.3330701

[29] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361

[30] Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. Management Science, 9(3), 458–467. https://doi.org/10.1287/mnsc.9.3.458

[31] Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. npj Digital Medicine, 3, 17. https://doi.org/10.1038/s41746-020-0221-y

[32] Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. Journal of Advanced Nursing, 32(4), 1008–1015. https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x

[33] Sendelbach, S., & Funk, M. (2013). Alarm fatigue: A patient safety concern. AACN Advanced Critical Care, 24(4), 378–386. https://doi.org/10.1097/NCI.0b013e3182a903f9

[34] Hernández, G., Ospina-Tascón, G. A., Damiani, L. P., Estenssoro, E., Dubin, A., Hurtado, J., Friedman, G., Castro, R., Alegría, L., Teboul, J. L., Cecconi, M., & Bakker, J. (2019). Effect of a resuscitation strategy targeting peripheral perfusion status vs serum lactate levels on 28-day mortality among patients with septic shock: The ANDROMEDA-SHOCK randomized clinical trial. JAMA, 321(7), 654–664. https://doi.org/10.1001/jama.2019.0071

[35] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[36] Mebrahtu, T. F., Skyrme, S., Randell, R., Keenan, A. M., Bloor, K., Yang, H., Andre, D., Ledward, A., King, H., & Thompson, C. (2021). Effects of computerised clinical decision support systems (CDSS) on nursing and allied health professional performance and patient outcomes: A systematic review of experimental and observational studies. BMJ Open, 11(12), e053886. https://doi.org/10.1136/bmjopen-2021-053886