Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research Paper

Development and validation of a deep interpretable network for continuous acute kidney injury prediction in critically ill patients

Meicheng Yang [a,1], Songqiao Liu [b,c,1], Tong Hao [b,1], Caiyun Ma [a], Hui Chen [b], Yuwen Li [a], Changde Wu [b], Jianfeng Xie [b], Haibo Qiu [b], Jianqing Li [a,d,*], Yi Yang [b,**], Chengyu Liu [a,*]

[a] *The State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing, China*
[b] *Jiangsu Provincial Key Laboratory of Critical Care Medicine, Department of Critical Care Medicine, Zhongda Hospital, School of Medicine, Southeast University, Nanjing, China*
[c] *Department of Critical Care Medicine, Nanjing Lishui People's Hospital, Zhongda Hospital Lishui Branch, Nanjing, China*
[d] *School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China*

ARTICLE INFO

Keywords:
Acute kidney injury
Critical care
Predictive model
External validation
Model interpretation

ABSTRACT

Early detection of acute kidney injury (AKI) may provide a crucial window of opportunity to prevent further injury, which helps improve clinical outcomes. This study aimed to develop a deep interpretable network for continuously predicting the 24-hour AKI risk in real-time and evaluate its performance internally and externally in critically ill patients. A total of 21,163 patients' electronic health records sourced from Beth Israel Deaconess Medical Center (BIDMC) were first included in building the model. Two external validation populations included 3025 patients from the Philips eICU Research Institute and 2625 patients from Zhongda Hospital Southeast University. A total of 152 intelligently engineered predictors were extracted on an hourly basis. The prediction model referred to as DeepAKI was designed with the basic framework of squeeze-and-excitation networks with dilated causal convolution embedded. The integrated gradients method was utilized to explain the prediction model. When performed on the internal validation set (3175 [15 %] patients from BIDMC) and the two external validation sets, DeepAKI obtained the area under the curve of 0.799 (95 % CI 0.791–0.806), 0.763 (95 % CI 0.755–0.771) and 0.676 (95 % CI 0.668–0.684) for continuousAKI prediction, respectively. For model interpretability, clinically relevant important variables contributing to the model prediction were informed, and individual explanations along the timeline were explored to show how AKI risk arose. The potential threats to generalisability in deep learning-based models when deployed across health systems in real-world settings were analyzed.

1. Introduction

Acute kidney injury (AKI) is a serious clinical syndrome characterized by a rapid decline in renal function caused by a variety of etiology and pathological mechanisms. AKI affects 10–15 % of hospitalized patients [1] and 30–60 % of critically ill patients in the intensive care unit (ICU) [2]. It is also reported that AKI serves as an independent risk factor for all-cause in-hospital death for patients with coronavirus disease 2019 [3]. However, an audit in the UK found that nearly half of AKI diagnoses were recognized late or not at all [4]. While there are no specific interventions for the prevention of AKI, early detection of AKI and implemented early "care bundles" such as fluid status optimization and avoidance of nephrotoxins may be associated with improved outcomes [5].

Recent markedly increased amounts of electronic health record (EHR) data and advances in the field of artificial intelligence (AI) have led to the rapid growth in developing AI-based algorithms for AKI early prediction [6,7]. Previous studies have developed models for predicting AKI at 24 h after admission [8,9] or 48 h after admission [10,11]. However, the medical condition of critically ill patients can significantly change during the following ICU stays. Integrating dynamic high-dimensional healthcare data for continuously assessing the AKI risk of

* Corresponding authors at: No. 2 Sipailou, Nanjing, Jiangsu Province 210096, China.
** Correspondence to: Y. Yang, No. 87 Dingjiaqiao, Nanjing, Jiangsu Province 210009, China.
E-mail addresses: ljq@seu.edu.cn (J. Li), yiyiyang2004@163.com (Y. Yang), chengyu@seu.edu.cn (C. Liu).

[1] The authors contributed equally.

https://doi.org/10.1016/j.artmed.2024.102785
Received 19 September 2022; Received in revised form 5 October 2023; Accepted 21 January 2024
Available online 26 January 2024

patients is required [12]. Traditional machine learning models for continuous AKI prediction require a priori knowledge of the temporal change states of the patients and do not capture complex interactions between trends in clinical variables. Deep learning-based models that could automatically learn relevant trends from EHR have been rapidly raised, among which the state-of-the-art work was reported in Nature by Google Deepmind [13]. However, the model was developed on specific populations from the US Department of Veterans Affairs, lacking independent external validation, so whether it could generalize to new institutions remain unknown. It is reported that only 6 % of conducted studies performed external validation in research of applying AI in healthcare [14].

Clinical risk scores and simple logistic regression models are still commonly used in clinical practice due to understandability and transparency. However, their performance is limited. Despite the enhanced performance of complex AI technology, they have not been integrated in part due to poor interpretability. In this respect, deep neural networks have been criticized as being the black box of algorithms [15]. Therefore, understanding how the AI-based algorithms arrive at their predictions and getting insights into the exact changes in risk induced by certain characteristics of an individual patient are required for clinicians [16]. Methods such as layer-wise relevance propagation (LRP) [17] and integrated gradients (IG) [18] toward explaining deep neural networks have been proven successful on computer vision tasks, providing frameworks for explaining EHR data tasks especially acute critical illness prediction.

Table 1
Generated 152 features by category.

Category	Features
Demographics [10]	Age, Gender, Weight, Height, Chronic kidney disease, Diabetes, Chronic pulmonary disease, Congestive heart failure, Moderate/Severe liver disease, Current ICU length of stay
Vital signs [7]	Heart rate, Temperature, Systolic blood pressure, Mean arterial pressure, Diastolic blood pressure, Respiration rate, SpO_2
Laboratory values [27]	pH, pO_2, pCO_2, FiO_2, BaseExcess, Lactate, Glucose, Hematocrit, Hemoglobin, White blood cells, Platelet, Albumin, Aniongap, Bicarbonate, Blood urea nitrogen, Creatinine, Calcium, Chloride, Sodium, Potassium, International normalized ratio, Prothrombin time, Alanine aminotransferase, Alkaline phosphatase, Aspartate aminotransferase, Total bilirubin, Total 12-h urine output
Empiric features [5]	PO_2/FiO_2 ratio, Blood Urea Nitrogen (BUN)/SCr ratio, Body Mass Index (BMI), lowest SCr value in the last 48 h, total-12 h-urine output/weight/12 h (UO_12h_Rt)
Informative missingness features [34]	Binary indicator to distinguish between the missing value and an actual clinical event of each vital sign and laboratory values
Trend features [34]	The difference between the current record and the previous value of each vital sign and laboratory values
Statistics features [35]	For vital signs, statistics (maximum, minimum, median, standard deviation [SD], and differential SD) in a 24-h sliding window were counted

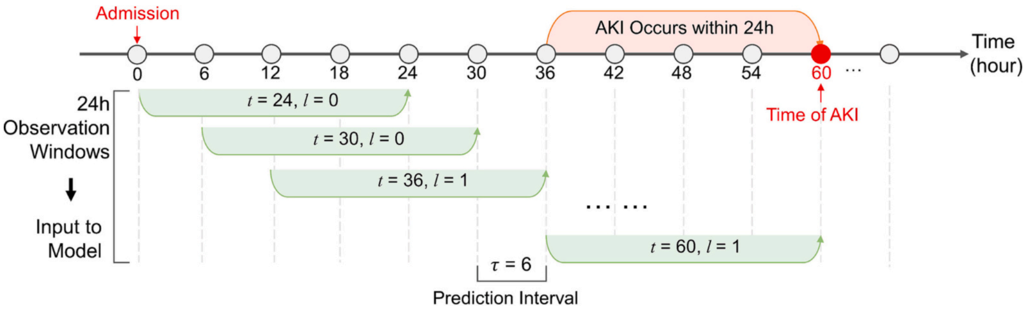

Fig. 1. Illustration of the discrete-time analyses and prediction task. The research target was to continuously calculate the risk of occurring AKI in the next 24 h at a regular time interval τ before the onset of AKI. The prediction label l at each prediction time t was considered positive if AKI onset occurred within 24 h; otherwise negative.

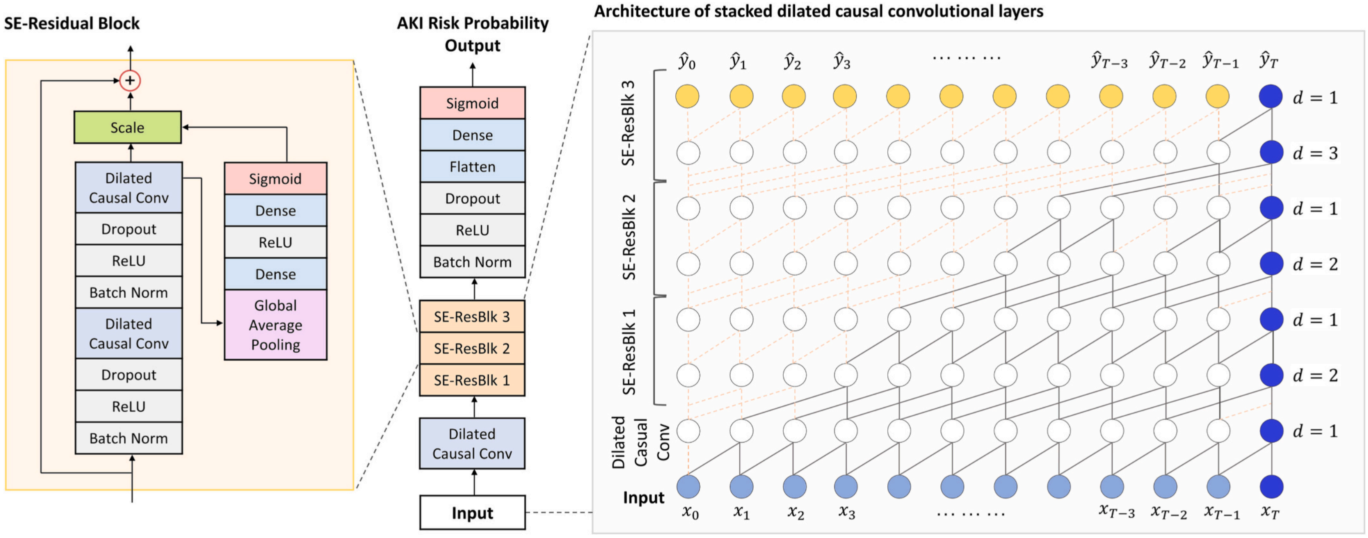

Fig. 2. The proposed deep neural network architecture. SE = squeeze-and-excitation, Conv = convolution, ReLU = rectified linear unit, Batch Norm = batch normalization, AKI = Acute kidney injury.

Table 2
Baseline patient characteristics in the development and validation sets.

Parameters[a]	Development A (N = 17,988)	Internal validation A (N = 3175)	External validation B (N = 3025)	External validation C (N = 2625)
Age (years)	65 (53–76)	64 (53–76)	63 (51–74)***	65 (52–76)
Male	10,283 (57.2)	1887 (59.4)*	1784 (59.0)	1591 (60.6)***
Height (cm)	170 (163–178)	170 (163–178)	170 (162–177)*	168 (160–172)***
Weight (kg)	75 (63–89)	76 (65–89)**	77 (64–93)***	65 (59–70)***
Comorbidity				
Chronic kidney disease	2325 (12.9)	376 (11.8)	193 (6.4)***	91 (3.5)***
Diabetes	4734 (26.3)	884 (27.8)	817 (27.0)	604 (23.0)***
Chronic pulmonary disease	4449 (24.7)	798 (25.1)	562 (18.6)***	225 (8.6)***
Congestive heart failure	4345 (24.2)	764 (24.1)	498 (16.5)***	189 (7.2)***
Moderate/severe liver disease	592 (3.3)	99 (3.1)	39 (1.3)***	177 (6.7)***
Admission SOFA score	4 (2–6)	4 (2–6)	3 (1–5)***	5 (3–7)***
Admission creatinine (mg/dL)	0.9 (0.7–1.2)	0.9 (0.7–1.3)	1 (0.8–1.5)***	0.8 (0.6–1.1)***
First day urine output (mL)	2224 (1600-3035)	2240 (1620-3115)	2175 (1500-3100)	2890 (2108-3913)***
ICU length of stays (days)	2.3 (1.6–4.0)	2.3 (1.6–4.0)	2.6 (1.8–4.2)***	7.1 (3.1–14.8)***
Any AKI stage	6628 (36.8)	1177 (37.1)	1052 (34.8)*	1279 (48.7)***
Hospital mortality	1125 (6.3)	196 (6.2)	181 (6.0)	166 (6.3)
Non-AKI	382 (3.4)	58 (2.9)	67 (3.4)	34 (2.5)**
Any AKI stage	743 (11.2)	138 (11.7)	114 (10.8)	132 (10.3)*

*p value between 0.01 and 0.05, **p value between 0.001 and 0.01; ***p value <0.001. Statistical analyses are conducted to describe the differences between the patients sourced from the development set and each validation set.
A = Medical Information Mart for Intensive Care database, B = eICU Collaborative Research Database, C = Chinese Database in Intensive Care, SOFA = Sequential Organ Failure Assessment.
[a] Data is in count (%) or median (Interquartile range, IQR).

In this study, we first developed a deep interpretable network for predicting AKI risk within 24 h for critically ill patients on a large ICU database. The proposed model aimed to help continuously detect physiological changes in patients, alert caregivers about patients at high risk of AKI and provide interpretable information for active treatments. We then compared the performance of this proposed model with three commonly used AI models on AKI prediction. After that, we validated all the developed AI models on two independent healthcare systems to explore the potential threats to the performance of AI-based algorithms when used in real-world settings.

2. Methods

2.1. Data sources

Data used was sourced from three distinct databases with different EHR systems: Medical Information Mart for Intensive Care database-IV [19] (MIMIC-IV, Metavision system), eICU Collaborative Research Database [20] (eICU-CRD, Philips eICU system) and a local Chinese Database in Intensive Care (CDIC, Wiicare system). MIMIC-IV captured de-identified health information for 76,540 ICU stays admitted to the ICUs at Beth Israel Medical Center (BIMC) between 2008 and 2019; eICU-CRD collated a multi-center dataset throughout the United States comprising 200,859 ICU admissions from 2014 to 2015. CDIC collected data from 6262 patients admitted to the Department of Critical Care Medicine, Zhongda Hospital Southeast University, China, from January 2018 to March 2021.

The open-source databases MIMIC-IV and eICU-CRD have received ethical approval from the Institutional Review Boards (IRBs) at BIDMC and Massachusetts Institute of Technology, CDIC database was approved by the IRB of Zhongda Hospital Southeast University on December 31, 2021 (2021ZDSYLL346–P01), conducted according to the principles outlined by the Helsinki Declaration, and a waiver for the requirement for informed consent was included in the IRB approval as all protected health information was de-identified.

2.2. Study population

We extracted data comprising all hospitalized adult patients who were first admitted to ICU for at least one day with at least an average of one urine output record every six hours and two serum creatinine (SCr) records during ICU admission. Patients were excluded if they had undergone any dialysis procedure or were diagnosed with the end-stage renal disease before the ICU visit. Patients who developed AKI or required renal replacement therapy during the period of 24 h before entering the ICU and 24 h after ICU admission were also excluded. AKI was defined according to the Kidney Disease: Improving Global Outcomes (KDIGO) criteria [21]. KDIGO accepts three definitions of AKI: [1] an increase in SCr of 0.3 mg/dL within 48 h; [2] an increase in SCr of 1.5 times the baseline creatinine, which is known or presumed to have occurred within the previous 7 days; [3] or a urine output of <0.5 mL/kg/h over 6 h. To obtain the baseline SCr, for patients admitted to the ICU, we also extracted all SCr values from laboratory events in the hospital system prior to ICU admission, so that the lowest value in the previous 7 days since the ICU admission could be calculated. However, if no SCr was recorded before ICU admission, the first SCr after ICU admission was used as the baseline SCr. In addition, the patient was required to be in the ICU for at least 6 h to achieve efficient urine output, and then we start using urine output to stage AKI. After that, AKI stages would be calculated at the time a clinical measurement of SCr or urine output was available.

2.3. Data preprocessing and feature generation

A total of 44 commonly available input variables were used in the model across various EHR systems, including demographics (age, gender, weight, height, comorbidity, current ICU length of stay), vital signs, and laboratory data (see Supplementary Table 1). Urine output information was integrated into 12 h (total 12 h urine output) as described by Koyner et al. [22]. All variables were time-ordered, and each ICU admission was represented by a sequence of clinical events. The event sequence was condensed into an hourly time window, and multiple events occurring within the same one-hour period were summarized as the average numerical value. We included all the time points for both patients with AKI and non-AKI during the length of stay before the first AKI occurrence or ICU discharge. Patient records were truncated to four weeks if they had an ICU length of stay for >28 days. Next, we employed one-hot encoding for the representation of categorical variables. All numerical features were standardized by removing the mean and scaling to unit variance after eliminating the outliers beyond the 1st and 99th percentile. For missing values during the event sequence, we carried forward the earlier available observation for imputation. If there was no available observation, we used the median values for the training data to fill in the remaining missing values. To avoid information leakage, the preprocessing operations were derived

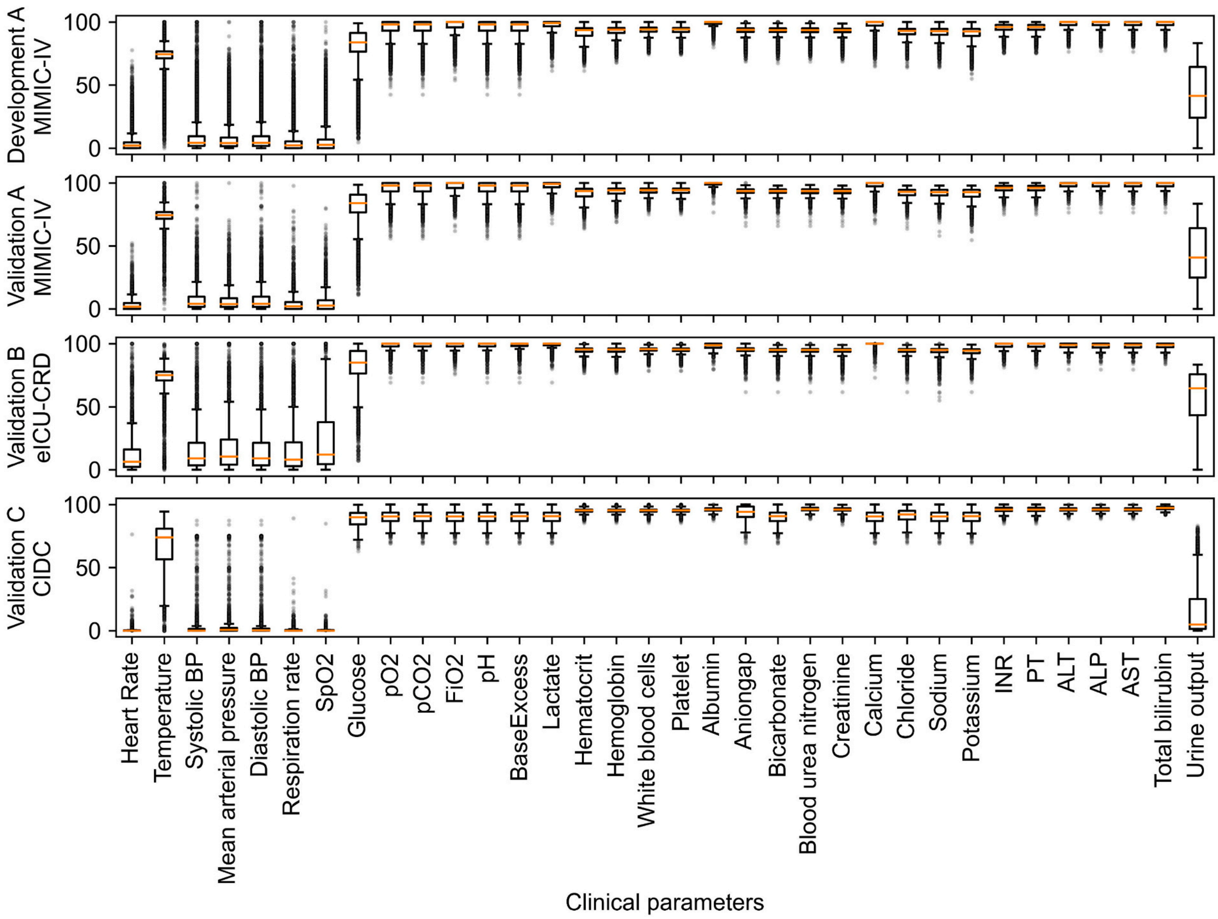

Fig. 3. Box plot for densities of missing data for the clinical parameters across sets. One dot indicates the proportion of missing data events in hours accounts for the total length of ICU stay. A = Medical Information Mart for Intensive Care database, B = eICU Collaborative Research Database, C = Chinese Database in Intensive Care, BP = blood pressure.

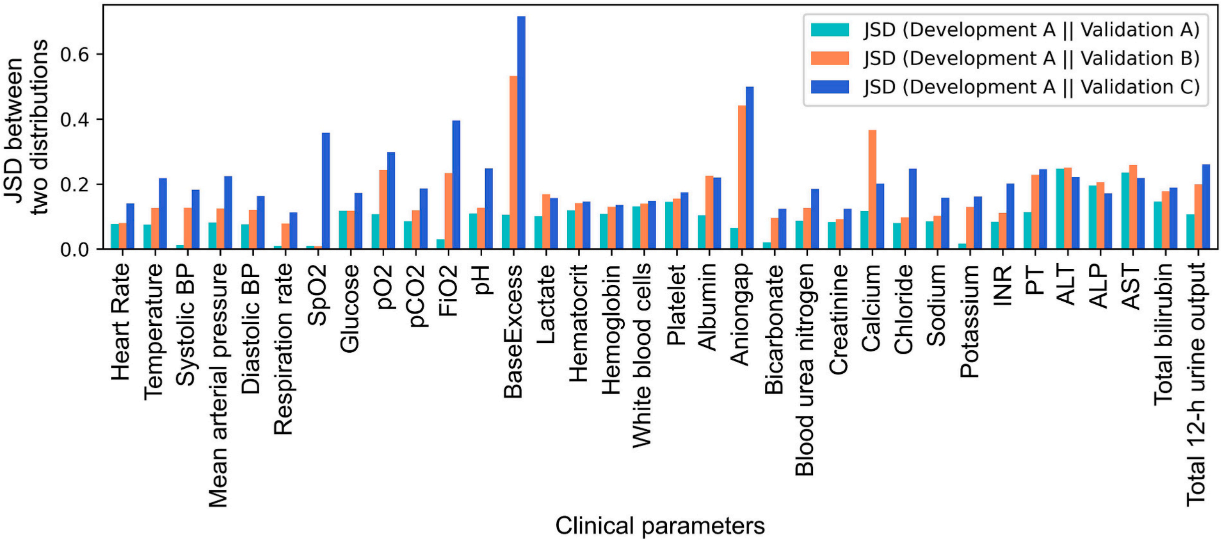

Fig. 4. Jensen-Shannon divergence (JSD) between two probability distributions of clinical parameters in the development set and each validation set. A = Medical Information Mart for Intensive Care database, B = eICU Collaborative Research Database, C = Chinese Database in Intensive Care, BP = blood pressure.

from the training data and applied to other datasets.

Apart from the 44 initial clinical variables, we also calculated another five features that were relevant for routine clinical practice, including PO_2/FiO_2 ratio, Blood Urea Nitrogen (BUN)/SCr ratio, Body Mass Index (BMI), lowest SCr value in the last 48 h, and total-12 h-urine output/weight/12 h (UO_12h_Rt). For the 34 vital signs and lab variables, we associated each with one binary indicator variable to enable the model to distinguish between the filled values and clinically

Table 3
Any AKI stage prediction performance for deepAKI and other models evaluated on internal and external validation populations.

Validation population	Number of patients and 24-h observation windows	Metrics[a] (95 % CI)	Proposed DeepAKI Neural Network	Long short term memory	Gradient boosting decision tree	Logistic regression
Internal validation A (MIMIC-IV)	NoW = 20,271 NoW-P = 4,327 (21.3 %)	AUC[b]	0.799 (0.791–0.806)	0.786 (0.778–0.794)	0.783 (0.775–0.790)	0.759 (0.752–0.767)
		SPC[b]	70.1 (68.4–71.7)	68.0 (66.3–69.8)	66.7 (64.9–68.3)	63.3 (61.8–64.7)
		PPV[b]	40.4 (39.1–41.8)	38.9 (37.7–40.3)	37.9 (36.7–39.1)	35.7 (34.8–36.6)
		NPV[b]	91.1 (91.0–91.4)	91.0 (90.7–91.1)	90.8 (90.5–91.0)	90.3 (90.1–90.5)
	NoP = 3175	SEN[c]	97.1 (96.4–97.9)	94.8 (93.8–95.8)	96.9 (96.3–97.7)	95.3 (94.4–96.3)
	NoP-P = 1,177 (37.1 %)	SPC[c]	42.1 (39.8–44.3)	38.8 (36.8–41.3)	37.7 (35.2–40.0)	31.5 (29.5–33.6)
External validation B (eICU-CRD)	NoW = 23,195 NoW-P = 4,145 (17.9 %)	AUC[b]	0.763 (0.755–0.771)	0.750 (0.741–0.759)	0.739 (0.731–0.748)	0.712 (0.704–0.721)
		SPC[b]	62.0 (60.0–63.7)	60.5 (58.2–62.9)	56.7 (55.2–59.2)	54.6 (52.8–57.1)
		PPV[b]	30.1 (29.0–31.0)	29.2 (28.1–30.6)	27.5 (26.7–28.5)	26.4 (25.7–27.5)
		NPV[b]	91.9 (91.7–92.1)	91.7 (91.5–92.0)	91.3 (91.0–91.6)	90.9 (90.7–91.3)
	NoP = 3025	SEN[c]	96.7 (95.9–97.5)	95.7 (94.7–97.0)	97.1 (96.2–97.9)	95.5 (94.5–96.6)
	NoP-P = 1,052 (34.8 %)	SPC[c]	34.2 (31.8–36.4)	32.8 (29.5–35.8)	29.1 (27.4–31.7)	24.6 (22.7–26.4)
External validation C (CDIC)	NoW = 52,938 NoW-P = 5,847 (11.0 %)	AUC[b]	0.676 (0.668–0.684)	0.655 (0.648–0.662)	0.654 (0.647–0.662)	0.660 (0.652–0.668)
		SPC[b]	45.4 (43.7–47.6)	44.4 (42.6–46.0)	42.3 (41.1–44.3)	45.6 (44.0–47.3)
		PPV[b]	14.6 (14.2–15.1)	14.3 (14.0–14.7)	13.9 (13.7–14.3)	14.6 (14.3–15.0)
		NPV[b]	93.6 (93.4–93.9)	93.4 (93.2–93.7)	93.2 (93.0–93.5)	93.6 (93.4–93.8)
	NoP = 2625	SEN[c]	98.7 (98.2–99.1)	98.4 (97.9–99.0)	99.5 (99.0–99.8)	98.7 (98.3–99.3)
	NoP-P = 1,279 (48.7 %)	SPC[c]	7.9 (6.4–9.3)	5.4 (4.3–6.5)	3.7 (2.9–4.7)	6.7 (5.6–7.8)

NoP = Number of Patients, NoP-P = Number of Positive Patients, NoW = Number of 24-h Observation Windows, NoW-P = Number of Positive Observation Windows, MIMIC = Medical Information Mart for Intensive Care database, eICU-CRD = eICU Collaborative Research Database, CDIC = Chinese Database in Intensive Care, AKI = Acute kidney injury, SEN = Sensitivity, SPC = Specificity, PPV = Positive predictive value, NPV = Negative predictive value, AUC = the Area Under the Curve.

[a] Threshold-based performance metrics were calculated and expressed as a percentage (%) when setting the sensitivity at 75 % in window-wise.

[b] Window-wise. Predictions were made when input 24-h observation windows in a six-hour interval.

[c] Patient-wise. Patients would be referred to as positive cases (AKI patients) if there was at least one prediction of the 24-h observation window during their ICU stay whose probability was higher than the threshold and negative cases if not.

measured values. A value of 1 is assigned when there is a clinical record at a certain time point, and a value of 0 is assigned when there is no record, indicating a missing value [13]. This operation resulted in 34 features. In addition, we also calculated the difference between the current clinical measurement and the previous clinical record as a form of simple trend features, thus resulting in another 34 features. For accessed vital signs, statistics (maximum, minimum, median, standard deviation [SD], and differential SD) in a 24-h sliding window were counted (a total of 35 features). Thus, we obtained a total of 152 features (for the full list, see Table 1).

2.4. Model development

The prediction target in this study was to continuously calculate the risk of occurring AKI in the next 24 h at a regular time interval τ before the onset of AKI, as illustrated in Fig. 1. We chose $\tau = 6$ h with suggestions from clinicians to avoid excessive alarms, which may result in alert fatigue. The observation window length was set as 24 h to capture as much information from laboratory tests as possible referred to the study by Song et al. [23]. We split the time series into the fixed 24-h window in a 6-h step after the first day, which means that we could make a prediction every 6 h based on the previous 24-h observed data (i.e., 24-h observation window) starting from the first day after ICU admission.

The prediction label at each prediction time t is a binary variable that is positive if AKI occurs within a 24-h future time horizon. If no AKI state was recorded within the future time horizon, the label was treated as negative. Predictor variables from the prior 24-h observation window preceding each prediction in a six-hour interval were used to train our model.

2.4.1. Neural network architecture

A deep neural network referred to as DeepAKI was developed that operated sequentially over the EHR for critically ill patients. The algorithm was designed with the basic framework of Squeeze-and-Excitation Networks (SENet) [24] with dilated causal convolution [25] instead of traditional one-dimensional convolution. SENet is a variation of a Resnet that has been demonstrated to be highly capable of performing visual or sequential tasks [24]. Fig. 2 gives a schematic view of our model. The model takes the observation sequence from patients as input and outputs the probability of AKI occurring in the next 24 h. Causal convolutions in the network premise that there can be no leakage from the future into the past, i.e., the key constraint for an output $\widehat{y}_t$ at time t is only dependent on elements from inputs $x_0, x_1, \ldots, x_t$ in the previous layer. We also used dilated convolutions to achieve a long effective history size by skipping input values with a certain step (dilation factor). This allows a better knowledge learning from both the most recent information and the much earlier states for accessing whether a patient would develop AKI.

The network consists of a dilated causal convolutional layer followed by three SE-Residual blocks with two dilated causal convolutional layers per block. Stacked dilated convolutions enable the network to have large receptive fields with just a few layers without greatly increasing computational costs. As displayed in Fig. 2, dilation factors set as $d = [1, 2, 1, 2, 1, 3, 1]$ with kernel size = 2 in this architecture enable output at the top level to yield a maximum receptive field of 12 h. The outputs of each convolutional layer with 64 filters are transformed using batch normalization [26] and fed into a rectified linear unit activation [27]. Spatial dropout [28] with a rate of 0.2 is added for regularization. In each block, the Squeeze-and-Excitation unit that adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels is integrated before the residual block is added to shortcut connections [24]. Outputs from the final spatial dropout are flattened to a single vector that is used as input to a final dense layer, followed by a sigmoid activation function. The output from the sigmoid activation is the probability of AKI risk in the next 24 h during ICU.

The binary cross-entropy loss function is minimized using the Adam optimizer [29] with a mini-batch size of 200. We initialized the learning rate to 0.001 and reduced it by a factor of five if the validation loss

Table 4
Model performance of the area under the curve across different clinical subgroups.

Subgroup name	Internal validation A (N = 3175)	External validation B (N = 3025)	External validation C (N = 2625)
Age (years)			
18–45	0.793 (0.769–0.817)	0.720 (0.697–0.743)	0.627 (0.605–0.647)
45–65	0.784 (0.771–0.796)	0.772 (0.757–0.786)	0.687 (0.674–0.699)
65–85	0.797 (0.786–0.809)	0.763 (0.751–0.775)	0.694 (0.682–0.705)
> 85	0.753 (0.723–0.784)	0.756 (0.725–0.786)	0.684 (0.656–0.713)
BMI (kg/m^2)			
< 18.5	0.744 (0.674–0.812)	0.727 (0.687–0.766)	0.616 (0.583–0.647)
18.5–25	0.796 (0.782–0.809)	0.773 (0.758–0.786)	0.668 (0.659–0.678)
25–30	0.782 (0.769–0.795)	0.755 (0.738–0.770)	0.668 (0.651–0.683)
> 30	0.806 (0.791–0.819)	0.758 (0.743–0.772)	0.672 (0.642–0.701)
Gender			
Female	0.794 (0.783–0.806)	0.757 (0.743–0.771)	0.695 (0.683–0.708)
Male	0.801 (0.791–0.811)	0.767 (0.756–0.777)	0.660 (0.652–0.670)
At risk groups			
CKD	0.772 (0.749–0.794)	0.707 (0.672–0.739)	0.765 (0.731–0.799)
Diabetes	0.809 (0.796–0.823)	0.775 (0.757–0.791)	0.727 (0.712–0.741)
Trauma	0.793 (0.774–0.813)	0.772 (0.741–0.801)	0.652 (0.634–0.670)
Sepsis	0.794 (0.783–0.805)	0.779 (0.756–0.802)	0.646 (0.632–0.660)
Cardiac surgery	0.773 (0.749–0.798)	0.793 (0.752–0.832)	0.802 (0.674–0.903)

A = Medical Information Mart for Intensive Care database, B = eICU Collaborative Research Database, C = Chinese Database in Intensive Care, AKI = Acute kidney injury, BMI = Body mass index, CKD = Chronic kidney disease.

stopped improving for five consecutive epochs. The neural network weights are initialized as described by He et al. [30]. The training runs for 100 epochs, with the final model being the one with the best validation results during the optimization process. Early stopping [31] and both L1 and L2 regularization are used to avoid overfitting. We applied a class weight for weighting the loss function to handle class imbalances.

2.4.2. Explanation module

There were three commonly used methods for explaining deep neural networks, occlusion analysis based on Shapley values, IG based on Taylor expansions, and LRP based on deep Taylor decompositions. Compared to occlusion analysis, IG produces finer point-wise explanations. In addition, the typical method, DeepSHAP [32], assumes that the input features are independent of each other and uses the linear composition rule in deep models, which could limit its usefulness in capturing the relationships between features and weaken the non-linear characteristics of deep neural networks. IG is also widely applicable to neural networks with complex structures and can be easily implemented in state-of-the-art deep learning frameworks such as PyTorch or TensorFlow, while LRP has a stronger requirement on the model structure. Therefore, we chose the IG method to interpret how the network works over the inputs and make the AKI prediction. For further reading on the methods and applications of explaining deep neural networks, see the review by Samek et al. [33]. IG highlights the feature which has the steepest local slope with respect to the output. Suppose the proposed network function f and input v, IG assigns an attribution value φ_i to the i-th feature by accumulating gradients interpolated in a way between the baseline b which represents the “missingness” of feature input and the specific input. The way can be represented by a path function $\gamma(\alpha)$ from the baseline b to the input v, where $\alpha \in [0,1]$, and $\gamma(0) = b, \gamma(1) = v$. A straight-line path was specified in IG, i.e., $\gamma(\alpha) = b + \alpha(v - b)$. Therefore, we derive that:

$$\varphi_i(f, v, b) = (v_i - b_i) \times \int_{\alpha=0}^{1} \frac{\delta f(b + \alpha(v - b))}{\delta v_i} d\alpha$$

Therefore, the attribution (importance score) of a specific feature can be accessed when computing the change output of the network starting with the baseline to the current value by integrating over a path and averting the problem of the local gradients in neural networks being saturated. However, the way to choose the baseline remains a problem. To avoid choosing a specific baseline, we average over multiple baselines to follow the expected gradients proposed by Erion et al. [34]. Assumpt that given a baseline distribution D, φ_i is redefined as the formula:

$$\varphi_i(f, v; D) = \int_b \varphi_i(f, v, b) \times p_D(b) db$$

where p_D represents the density function. Suppose α obeys the uniform distribution U between 0 and 1, expected gradients reformulate the integrals above as expectations:

$$\varphi_i(f, v; D) = \mathbb{E}_{b \sim D, \alpha \sim U(0,1)} \left[(v_i - b_i) \times \frac{\delta f(b + \alpha(v - b))}{\delta v_i} \right]$$

In this paper, to compute feature attribution values in practice, we simply use k samples from the given populations of the development set as random samples from D and obtain the formula as follows:

$$\varphi_i(f, v, k; D) = \frac{1}{k} \sum_{j=1}^{k} \left(v_i - b_i^j\right) \times \frac{\delta f\left(b^j + \alpha^j \left(v - b^j\right)\right)}{\delta v_i}$$

We apply two aspects to visualize the explanations, including individual interpretability and global understanding. By computing IG, individual interpretability assigns an attribution value to each feature for each 24-h observation window at every prediction time point during ICU stay for a single patient. It explains how the model outputs probability and reminds us of relevant risk factors. Global understanding tells the most relevant clinical parameters for predicting AKI after learning from all given populations.

2.5. Experiments setup

Patients in the MIMIC-IV database were randomly divided into a development set (85 %, renamed Development A, consisting of a 70 % training dataset and a 15 % hyperparameter tuning dataset) for training and an internal validation set (15 %, Validation A) used to provide an unbiased evaluation. To compare the performance of the proposed approach, models including an attention-based Long Short Term Memory (LSTM) neural network [8,35], a Gradient Boosting Decision Tree (GBDT) model [22], and a discrete-time logistic regression (LR) approach [36] that have commonly been used for AKI prediction were also trained. All the models' hyperparameters were optimized using the grid search method performed on the 15 % hyperparameter tuning dataset (Supplementary Table 2). After that, we validated all the constructed AI models on two independent healthcare systems. In order to keep the size of patients and the total of predictions in hours as consistent as possible with the internal validation populations and make a horizontal comparison for performance, randomly 15 % of patients from eICU-CRD were used for one external validation (Validation B). The entire CDIC cohorts were used for another external validation (Validation C). The data for a single patient was assigned to one independent set to avoid information leakage.

To describe the differences of clinical parameters between the patients sourced from the development set and each validation set,

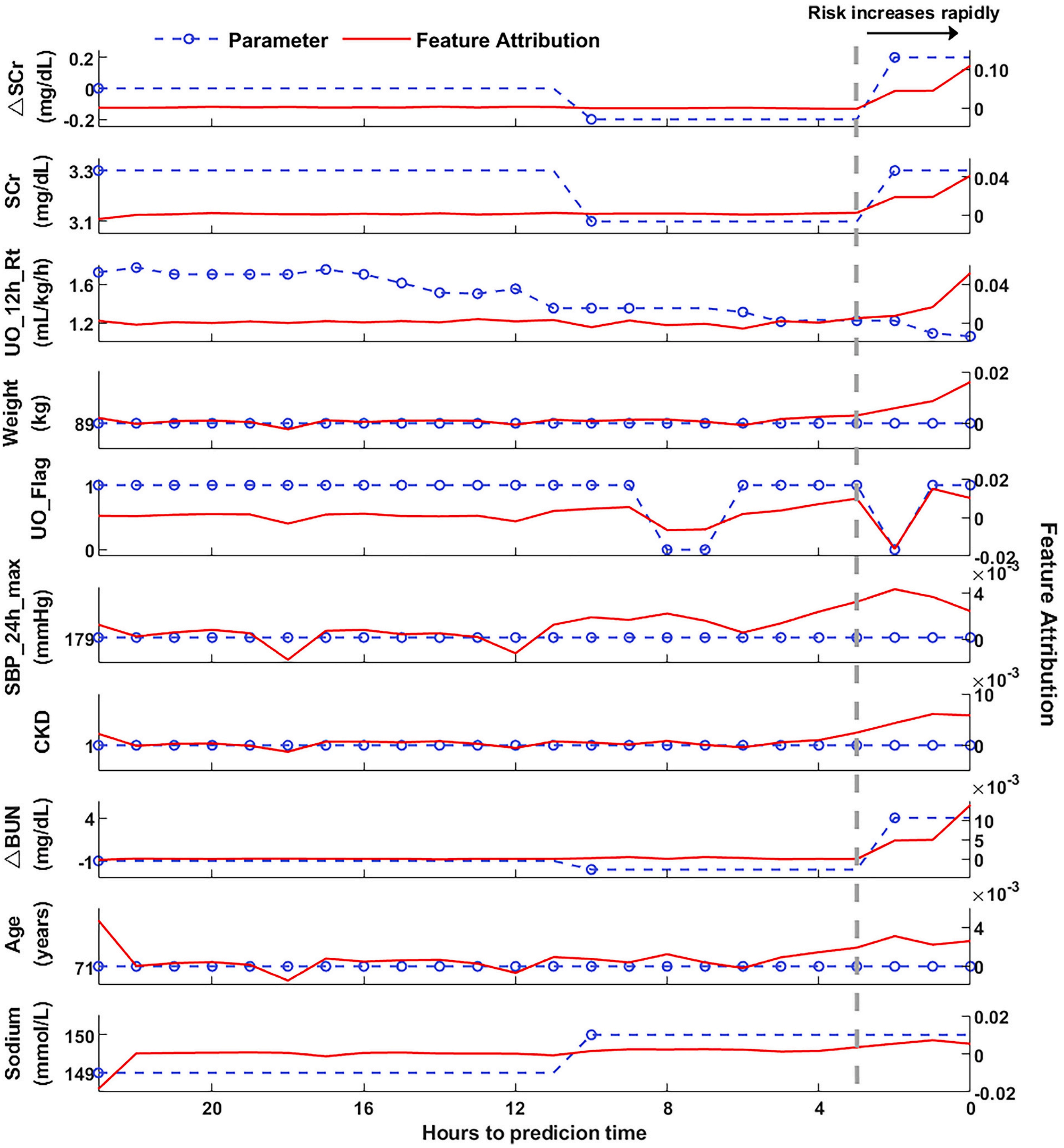

Fig. 5. An illustrated example of individual explanation at the prediction time 24 h before the stage 1 AKI onset for one male patient who was 71 years old with complications of chronic kidney disease and diabetes. The time point of zero above represents the prediction time, and DeepAKI warns of an AKI risk of 71.6 %. The parameter with a feature attribution value above zero pushes the AKI risk higher, otherwise lower. The Grey dotted line refers to the time of risk starts to increase rapidly. Dots that are recorded on the blue dotted line along the timeline of a parameter indicate the actual clinical events rather than imputation values. AKI = acute kidney injury, SCr = serum creatinine, ΔSCr = change in SCr, UO_12h_Rt = total-12 h-urine output/weight/12 h, UO_Flag = binary indicator to distinguish between the imputation value and actual measurement of urine output, SBP_24h_max = maximum of systolic blood pressure in the last 24 h, CKD = chronic kidney disease, ΔBUN = change in blood urea nitrogen. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

statistical analyses were conducted by using a chi-square test for categorical variables and the two-sided Wilcoxon rank-sum test for all continuous variables. We also applied the Jensen-Shannon-Divergence (JSD) to assess the similarity of distributions for each clinical measurement between the development set and validation set. JSD is an asymmetric metric in probability theory and statistics that can measure the relative entropy or difference in information represented by two distributions. It can be used to calculate the distance between two data distributions, providing insight into how different the two distributions are from each other. The JSD value falls between 0 and 1, the closer to 0, the more similar the two distributions are.

The discrimination performance of the model was mainly assessed by the area under the curve (AUC) after making all predictions of the 24-h observation window (window-wise) from the validation populations. The risk probability threshold was obtained when setting the sensitivity at 75 % in window-wise, so we can report the threshold-based performance metrics of specificity, positive predictive value (PPV), and negative predictive value (NPV), respectively. In addition, in order to

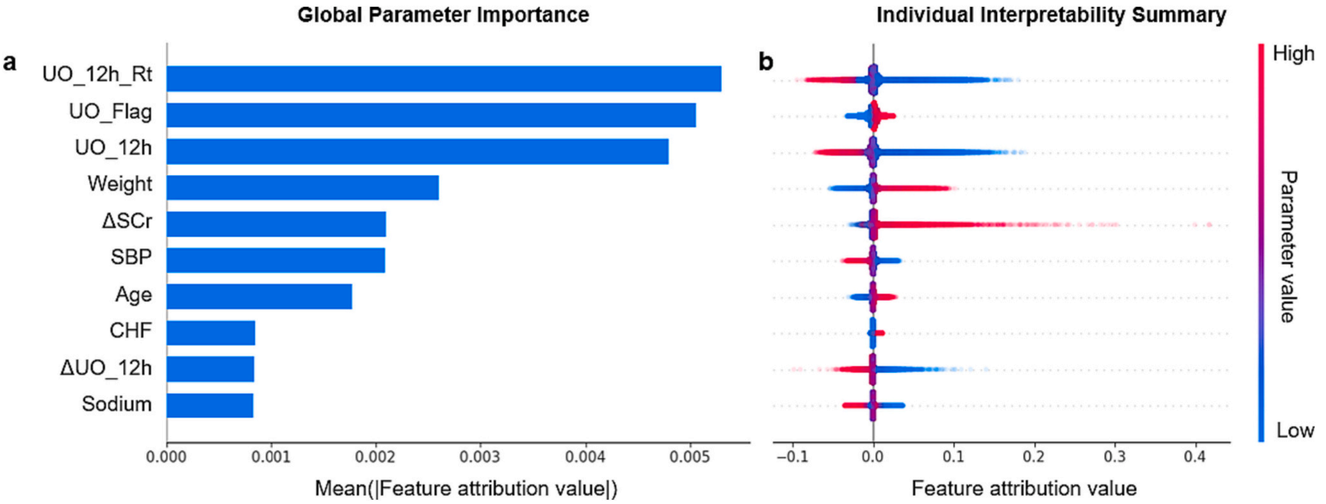

Fig. 6. Summary of feature attribution for the employed clinical parameters. a, global parameter importance of the top 10 features. b, beeswarm plots show parameter attribution values across patients for the top 10 features, where each dot indicates the attribution value for a one-hour sample. When summarizing the interpretability, the temporal relevance variations are simplified and ignored, treating all data points at different times equally. When multiple dots fall on the same *x* position, they are stacked to show density. Parameters with positive attribution values push the AKI risk higher, while negative push the risk lower. Longtails indicate features are extremely important for specific patients. AKI = Acute kidney injury, UO_12h_Rt = total-12 h-urine output/weight/12 h, UO_Flag = binary indicator to distinguish between the missing value and actual measurement of urine output, UO_12h = total 12 h urine output, SCr = serum creatinine, ΔSCr = change in SCr, SBP = systolic blood pressure, CHF = congestive heart failure, ΔUO_12h = change in UO_12h, HR = heart rate, BUN = blood urea nitrogen.

know how many patients the model can successfully predict or detect, we treated the patients as positive cases (AKI patients) if there was at least one prediction of the 24-h observation window during their ICU stay whose probability was higher than the threshold and negative cases if not. Therefore, the AKI/non-AKI predictions for all patients could be identified patient-wise, so that the metrics of sensitivity and specificity in patient-wise can be reported [37]. We calculated the 95 % CIs for all the performance measures using bootstrapping (1000 stratified bootstrap replicates).

3. Results

3.1. Patients description

A total of 17,988 patients were enrolled in the development set A, with 6628 (36.8 %) of whom developed at least stage 1 AKI. The validation sets at A, B, and C comprised 3175 patients (1177 [37.1 %] any AKI), 3025 patients (1052 [34.8 %] any AKI), and 2625 patients (1279 [48.7 %] any AKI), respectively (Table 2). Hospital mortality of patients who were diagnosed with AKI was significantly higher than those who did not on all the validation populations ([10.3–11.7 %] vs [2.5–3.4 %]). Compared with the patients from validation A and B, populations from validation set C had significantly longer median (IQR) ICU lengths of stay (7.1 [3.1–14.8] vs 2.3 [1.6–4.0] vs 2.6 [1.8–4.2] days), and were diagnosed with AKI much later (3.7 [1.9–6.6] vs 1.6 [1.3–2.4] vs 1.7 [1.3–2.7] days). The proportion of missing data accounting for the total length of ICU stay in hours for each vital sign and laboratory values were computed (Fig. 3). Most vital signs were recorded on an hourly basis in most patient records, while most laboratory values were sampled on a daily basis. There were large differences in distributions for the majority of vital signs and laboratory values between the development set and external validation sets according to the statistical analysis (Supplementary Table 1) and JSD results indicated (Fig. 4).

3.2. Model performance

Of the four developed AI models for the AKI prediction in internal validation A, DeepAKI performed best with an AUC of 0.799 (95 % CI 0.791–0.806) compared with 0.759–0.786 for the other models. In external validation, DeepAKI still performed best with an AUC of 0.763 (95 % CI 0.755–0.771) in validation B and 0.676 (95 % CI 0.668–0.684) in validation C. In contrast, the AUC performance of the other models was 0.712–0.750 in B and 0.654–0.660 in C. After setting the sensitivity at 75 % (window-wise) to obtain the risk probability threshold, we got a specificity of 70.1 % in A, 62.0 % in B, and 45.4 % in C. The ratio of false to true alarms of any AKI episodes was approximately 1.5 (PPV 40.4 %) and 2.3 (PPV 30.1 %) when performed on validation A and B. However, the ratio of false to true alarms is about 5.8 (PPV 14.6 %), which was high on the independent set C. In addition, performance in patients' statistics (patient-wise) showed that DeepAKI identified the majority of the AKI patients in all three healthcare systems (sensitivity ≥96.7 %). Still, it falsely identified many non-AKI patients as AKI, especially on the external validation C (specificity = 7.9 %) (Table 3).

For AKI prediction at-risk groups, DeepAKI performed better in diabetics patients with AUC of 0.809 (95 % CI 0.796–0.823) for whom from validation set A, while it performed better for patients after cardiac surgery from both validation set B with AUC of 0.793 (95 % CI 0.752–0.832) and validation set C with AUC of 0.802 (95 % CI 0.674–0.903) (Table 4).

3.3. Model interpretation

We first illustrated the individual interpretability for a single patient at the prediction time point 24 h before the stage 1 AKI onset (Fig. 5). The top 10 relevant clinical parameters ranking by an average magnitude of feature attribution values contributed to stage 1 AKI with a risk score of 71.6 % were shown. Parameter values and the corresponding feature attribution values along the timeline in this 24-h observation window were displayed to reflect the real-time state change of the patient. As for global understanding, the top 10 important features that the predictive model contributed to any AKI prediction were summarized in Fig. 6. The visible interpretability summary of feature attribution across patients revealed that oliguria, older, overweight, and hypotension patients were consistent with higher AKI risk, as shown in the right column of the figure.

4. Discussion

In this retrospective study, we built a deep interpretable network called DeepAKI learning from the collected sequential EHR data for continuously predicting the risk of developing AKI in the next 24 h at six-hour intervals. Results indicated that DeepAKI showed superior performance compared with LSTM, GBDT, and LR on both the internal and external validation populations. The proposed DeepAKI could provide quantitative and explainable risk factors contributing to the model prediction for a single patient, which could help improve the practicality of clinical decision support. We included both an independent US-sourced database and a China-sourced dataset with larger population heterogeneity for external validation to simulate deploying predictive models to real-world settings. Performance deterioration was found when deployed all the predictive models to external validation sets, especially the Chinese Database in Intensive Care.

Previous studies on AKI early prediction that had a fixed prediction time [8–11] limited their clinical practicalities. Daily prediction might delay the judgment of AKI risk [23,38], while hourly prediction would result in repeated alerts causing alert fatigue [36]. In this study, the model was designed to generate predictions every six hours, helping decrease the overall possible number of alerts when effectively tracking the disease progression rate of AKI [13,39]. Results demonstrated that DeepAKI outperformed the commonly used models on the AKI continuous prediction. Even after removing the most important five features UO_12h_Rt, UO_Flag, UO_12h, Weight, and ΔSCr, the results still showed acceptable model performance. The AUC (95 % CI) for validation A, B, and C were 0.761 (0.752–0.769), 0.713 (0.704–0.721), and 0.635 (0.627–0.642), respectively, which showed the DeepAKI's robustness. In addition, to the best of our knowledge, no study has evaluated the performance of AI-based models for continuous AKI prediction to multiple-sourced databases, especially from another country. Although previous AKI prediction models have achieved an AUC ranging from 0.73 to 0.78 in internal validation studies, and 0.60–0.76 in external validation studies [22,23,36,40], populations were from the same region or using the same EHR system.

Our experimental results demonstrated the performance deterioration and drew our attention to the following potential threats to generalisability in AI-based models in healthcare. First, the population heterogeneity. For example, patients from healthcare system A were more likely to have chronic kidney disease (CKD), chronic pulmonary disease, and congestive heart failure, while patients from C seemed to have more liver disease. Higher admission SOFA scores and longer ICU stay for patients from C were also observed. Nevertheless, they were diagnosed with AKI much later due to the higher urine output during the first day, and lower admission SCr, indicating kidney disease progresses relatively slowly. Second, software diversity of EHR for data capture. There is a data transformation problem for mapping local concepts to common terminology, especially across different EHR systems. For example, comorbidities were extracted in the public database A and B based on the International Classification of Diseases codes. However, we matched the Chinese field to find the comorbidities in local set C due to the incomplete disease recording mechanisms, which might underestimate the number of patients with comorbidities. Third, differences in clinical practice between regions. Caregivers at the bedside from C recorded data more frequently, resulting in relatively few missing values for vital signs, as shown in Fig. 3. The problems above caused the deviation of the data distribution and led to the decline in the generalization performance of the AI algorithm.

Another strength was that we first used the IG method computing the feature attribution values to uncover the black box of the proposed deep neural network for acute critical illness prediction. Previous studies explained AI models in predicting AKI or other critical illness risks using Wald z-scores of covariates [36], gain [22], SHapley Additive exPlanation (SHAP) values [23,41–43], attention mechanism [8], and LRP [44]. However, most of them either captured the overall behaviors of the model lacking individual explanations or mainly explained in tree-based models. The proposed DeepAKI could not only inform clinically relevant important variables contributing to the model prediction (Fig. 6) but also help clinicians easily understand why a model is predicting a certain diagnosis for a single patient. As the example shown in Fig. 5, we could successfully predict the AKI risk 24 h before the stage 1 AKI onset and remind how the risk rose. For the dynamic parameters in this 24-h observation window, especially the last few hours, it can be seen that the increase of SCr from 3.1 mg/dL to 3.3 mg/dL, BUN increased by 4 mg/dL and the continuous decrease in urine output highlight the AKI risk. Followed by the static demographic variables such as weight, CKD, and age raise more attention during this period.

Demystifying the prediction model not only helped us understand the model better but also helped further analysis of the source of performance heterogeneity [23]. For example, the feature of binary indicator to distinguish between the missing value and actual measurement of urine output (UO_Flag) was important in predicting AKI learning from populations in healthcare system A. However, urine output was recorded almost every hour in healthcare system C, so UO_Flag could not provide enough information for alerting AKI, which resulted in model performance degradation. When we removed the type of Informative Missingness Features (total of 34 features) and retrained the model, results indicated that the AUC (95 % CI) of the validation C can be improved from 0.676 (0.668–0.684) to 0.692 (0.685–0.699), even though the AUC of validation A dropped from 0.799 (0.791–0.806) to 0.787 (0.779–0.794), and the AUC of validation B dropped from 0.763 (0.755–0.771) to 0.748 (0.740–0.757).

This study was subject to some limitations. First, the BIMC development cohort spans from 2008 to 2019, during which time there may have been changes in the identification of AKI and data collection in inpatient settings. We also did not consider the impact of drugs and treatments such as diuretics on the urine output, which could mislead the AKI label. In addition, the partial absence of urine output records and the use of the first SCr after ICU admission as the baseline in some patients may affect the accuracy of AKI staging. Second, we only included information from demographics, vital signs, and laboratory values which limited the model performance and were unable to dig out more potential factors that lead to AKI. Further studies are going to collect more information, such as management with drugs and intravenous fluids, pressors, etc. Third, although we successfully used the IG method to explain the deep learning model, prospective studies are needed to verify how to apply it in clinical decision-assisted systems to provide truly helpful information. Additionally, the IG method only demonstrated learned correlations that were already established knowledge in the clinical field. Future studies are required to use this technology to discover new correlations and capture potential interaction effects between features. Fourth, the relatively high false alarm rate still brings difficulties in applying the model in clinical practice. An effective false alarm handling mechanism needs to be proposed for continuous critical illness prediction tasks in further research. Lastly, the problem of model generalizability was not well solved in this study. Further study is going to collect more data from multiple hospital systems to improve the model's performance. Techniques from transfer learning [45] or federated learning [46] that might improve model generalizability by adjusting data distribution shifts or pretraining will also be explored in further research. Nevertheless, this study highlights that when deploying across hospital institutions, it is important to understand the heterogeneity and minimize differentiation between features as much as possible.

5. Conclusions

The proposed model DeepAKI that continuously predicts AKI achieved superior performance compared with three commonly used AI algorithms. When externally validated all the AI algorithms, the results got deteriorated, which drew our attention to the potential threats to the generalisability of AI-based models when deployed across health systems in real-world settings. The model interpretability proposed could help improve clinical understanding of AKI risk at the global and individual patient levels and explain the model deterioration for further improvement.

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.

Acknowledgments

This work was funded by the National Key Research and Development Program of China (2023YFC3603600), the National Natural Science Foundation of China (62171123, 62071241, 6590009703, 81971885 and 81971888), the Scientific Research Project of Jiangsu Health Committee (H2018048, BE2019749), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21_0088). The authors are grateful to Liwei Huang, Yufen Zhu, and other intensive care unit doctors for helping provide clinical knowledge in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102785.

References

[1] Zeng X, McMahon GM, Brunelli SM, et al. Incidence, outcomes, and comparisons across definitions of AKI in hospitalized individuals. Clin J Am Soc Nephrol 2014;9:12–20.
[2] Hoste EA, Kellum JA, Selby NM, et al. Global epidemiology and outcomes of acute kidney injury. Nat Rev Nephrol 2018;14:607–25.
[3] Nadim MK, Forni LG, Mehta RL, et al. COVID-19 associated acute kidney injury: consensus report of the 25th Acute Disease Quality Initiative (ADQI) Workgroup. Nat Rev Nephrol 2020;16:747–64.
[4] Joslin J, Wilson H, Zubli D, et al. Recognition and management of acute kidney injury in hospitalised patients can be partially improved with the use of a care. Clin Med 2015;15:431–6.
[5] Kolhe NV, Reilly T, Leung J, et al. A simple care bundle for use in acute kidney injury: a propensity score-matched cohort study. Nephrol Dial Transplant 2016;31:1846–54.
[6] Vlieger GD, Kashani K, Meyfroidt G. Artificial intelligence to guide management of acute kidney injury in the ICU: a narrative review. Curr Opin Crit Care 2020;26:563–73.
[7] Rank N, Pfahringer B, Kempfert J, et al. Deep-learning-based real-time prediction of acute kidney injury outperforms human predictive performance. NPJ Digit Med 2020;3:139.
[8] Chen Z, Chen M, Sun X, et al. Analysis of the impact of medical features and risk prediction of acute kidney injury for critical patients using temporal electronic health record data with attention-based neural network. Front Med 2021;8:658665.
[9] Kate RJ, Perez RM, Mazumdar D, et al. Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak 2016;16:39.
[10] Malhotra R, Kashani KB, Macedo E, et al. A risk prediction score for acute kidney injury in the intensive care unit. Nephrol Dial Transplant 2017;32:814–22.
[11] Parreco J, Soe-Lin H, Parks JJ, et al. Comparing machine learning algorithms for predicting acute kidney injury. Am Surg 2019;85:725–9.
[12] Kate RJ, Pearce N, Mazumdar D, et al. A continual prediction model for inpatient acute kidney injury. Comput Biol Med 2020;116:103580.
[13] Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572:116–9.
[14] Kim DW, Jang HY, Kim KW, et al. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019;20:405–10.
[15] Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20–3.
[16] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
[17] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 2015;10:e0130140.
[18] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, 2017, pp. 3319–3328.
[19] Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.
[20] Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018;5:180178.
[21] Kidney disease improving global outcomes acute kidney injury work group: KDIGO clinical practice guideline for acute kidney injury. Kidney Int 2012;Suppl 2:1–138.
[22] Koyner JL, Carey KA, Edelson DP, et al. The development of a machine learning inpatient acute kidney injury prediction model. Crit Care Med 2018;46:1070–7.
[23] Song X, Yu ASL, Kellum JA, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun 2020;11:5668.
[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City; 2018. p. 7132–41.
[25] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proceedings of the International Conference on Learning Representations. San Juan; 2016. p. 1–13.
[26] Ioffe S, Szegedy C: Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015, arXiv:150203167.
[27] Nair V, Hinton GE: rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. Haifa, 2010, pp. 807–814.
[28] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. JMLR 2014;15:1929–58.
[29] Kingma DP, Ba J: Adam: a method for stochastic optimization. 2017, arXiv:14126980.
[30] He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. 2015, arXiv:151203385.
[31] Caruana R, Lawrence S, Giles L: Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In: Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver, 2000, pp. 381–387.
[32] Lundberg SM, Lee S: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, 2017, pp. 4768–4777.
[33] Samek W, Montavon G, Lapuschkin S, et al. Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE 2021;109:247–78.
[34] Erion G, Janizek JD, Sturmfels P, et al. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nat Mach Intell 2021;3:620–31.
[35] Kaji DA, Zech JR, Kim JS, et al. An attention based deep learning model of clinical events in the intensive care unit. PLoS One 2019;14:e0211057.
[36] Simonov M, Ugwuowo U, Moreira E, et al. A simple real-time model for predicting acute kidney injury in hospitalized patients in the US: a descriptive modeling study. PLoS Med 2019;16:e1002861.
[37] Shashikumar SP, Wardi G, Malhotra A, et al. Artificial intelligence sepsis prediction algorithm learns to say “I don’t know”. NPJ Digit Med 2021;4:134.
[38] Kim K, Yang H, Yi J, et al. Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: external validation and model interpretation. J Med Internet Res 2021;23:e24120.
[39] Dong J, Feng T, Thapa-Chhetry B, et al. Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care. Crit Care 2021;25:288.
[40] Churpek MM, Carey KA, Edelson DP, et al. Internal and external validation of a machine learning risk score for acute kidney injury. JAMA Netw Open 2020;3:e2012892.
[41] Tseng P-Y, Chen Y-T, Wang C-H, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. Crit Care 2020;24:478.
[42] Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care 2019;23:112.
[43] Yang MC, Liu CY, Wang XY, et al. An explainable artificial intelligence predictor for early detection of sepsis. Crit Care Med 2020;48:e1091–6.
[44] Lauritsen SM, Kristensen M, Olsen MV, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 2020;11:3852.
[45] Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. Proc IEEE 2021;109:43–76.
[46] Dayan I, Roth R, Zhong A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 2021;27:1735–43.