

Development and external validation of a machine-learning model for classifying documented pressure injury status during ICU stays among patients receiving invasive mechanical ventilation: A MIMIC-IV and eICU study

Yang Zhao^a, Siji Chen^b, Chang Pang^a, Yao Tang^a, Muhua Zeng^c, Shuwu Lin^{a,*}

^a Department of General Practice, The Second Affiliated Hospital of Shenyang Medical College, Shenyang, Liaoning, 110002, China

^b Cardiac Surgery Center, Beijing Anzhen Hospital, Capital Medical University, Beijing Institute of Heart, Lung and Blood Vessel Diseases, Beijing, 100029, China

^c Jinzhou Medical University, Jinzhou, Liaoning, 121001, China

ARTICLE INFO

Keywords:

Pressure injury
Intensive care unit
Invasive mechanical ventilation
MIMIC-IV
eICU
Machine learning

ABSTRACT

Objective: To develop and externally validate a machine-learning model for classifying documented pressure injury status during hospitalization among intensive care unit (ICU) patients receiving invasive mechanical ventilation.

Methods: The development cohort included 1601 patients from MIMIC-IV version 3.1, including 390 with and 1211 without documented pressure injury. Mean values of continuous clinical variables recorded during the 24 h before ICU admission were used. Variables and patients with >30% missing data were excluded, and remaining missing values were imputed using random forest imputation. Features were selected using Boruta and least absolute shrinkage and selection operator (LASSO) regression. Six machine-learning models were compared using 10-fold cross-validation, and XGBoost was subsequently optimized using GridSearchCV with ROC-AUC as the optimization metric and a random seed of 42. SHapley Additive exPlanations (SHAP) were used for model interpretation. External validation was performed in 1576 eICU patients, including 221 with and 1355 without documented pressure injury; the first value recorded after ICU admission was used because continuous pre-ICU data were unavailable.

Results: XGBoost showed the best overall performance among the six candidate models. After optimization, the model achieved a 10-fold cross-validated area under the receiver operating characteristic curve (AUC) of 0.853 (95% CI, 0.765–0.941) in MIMIC-IV and an AUC of 0.703 (95% CI, 0.669–0.737) in the independent eICU cohort. The final model included sex, white blood cell count, blood glucose, hematocrit, blood urea nitrogen, and anion gap. SHAP analysis identified blood urea nitrogen and anion gap as the strongest contributors to model output.

Conclusion: The optimized XGBoost model demonstrated moderate external discrimination for classifying documented pressure injury status among ICU patients receiving invasive mechanical ventilation and may provide supplementary information for clinical assessment.

1. Introduction

Pressure injury is a common complication among adults admitted to the intensive care unit (ICU) [1]. Previous studies have reported an overall pressure injury prevalence of 26.6% among ICU patients and an incidence of ICU-acquired pressure injury of 16.2% [2]. Pressure injuries may cause pain, infection, prolonged hospital stay, and increased healthcare costs and are closely associated with poor outcomes and

increased mortality [3]. Pressure injury in ICU patients has a multifactorial etiology, with major contributing factors including immobility, septic shock, vasoactive drug use, sedation, and mechanical ventilation [4]. Patients receiving invasive mechanical ventilation are particularly vulnerable because of sedation and analgesia, restricted mobility, hemodynamic instability, and multiple comorbidities [5].

Pressure injury prevention has become an important clinical priority as the quality of ICU nursing care continues to improve. Although the

* Corresponding author. Department of General Practice, the Second Affiliated Hospital of Shenyang Medical College, China.

E-mail addresses: 15164222196@163.com (Y. Zhao), linshu1967@126.com (S. Lin).

<https://doi.org/10.1016/j.jtv.2026.101058>

Received 25 July 2026; Received in revised form 20 September 2026; Accepted 26 September 2026

Available online 28 September 2026

0965-206X/© 2026 The Authors. Published by Elsevier Ltd on behalf of Society of Tissue Viability. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Braden Scale is widely used and provides useful guidance for risk assessment, it is relatively subjective and time-consuming [6]. In contrast, classification models based on electronic health record data may help clinicians identify ICU patients with documented pressure injury during hospitalization [7], thereby supporting targeted assessment and preventive care. Because machine-learning methods can analyze high-dimensional data and identify nonlinear relationships [8], they have been increasingly applied to outcome prediction in critically ill patients [9]. Although machine-learning models have been developed for mechanically ventilated ICU patients, most have been developed and validated within a single dataset; external validation across independent databases remains limited [10].

Therefore, we used the MIMIC-IV database to develop a classification model for documented pressure injury status during hospitalization among ICU patients receiving invasive mechanical ventilation and externally validated the final model using the eICU database. Shapley additive explanations (SHAP) were used to interpret model output. The findings are intended to help ICU clinicians identify patients with a high model-predicted probability of documented pressure injury during hospitalization and to provide supplementary information for comprehensive clinical assessment.

2. Materials and methods

2.1. Study design and population

Data for this study were obtained from two large, publicly available critical care databases. Development data were derived from the Medical Information Mart for Intensive Care IV (MIMIC-IV), a large, deidentified critical care database established by Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA. MIMIC-IV contains data from more than 65,000 ICU admissions and more than 200,000 emergency department visits and integrates multidimensional clinical information, including demographic characteristics, vital signs, laboratory measurements, medication records, and therapeutic interventions [11].

External validation data were obtained from the eICU Collaborative Research Database, which was jointly developed by Philips Healthcare and the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT). The database contains more than 200,000 ICU admissions and includes deidentified clinical data collected from ICUs in 208 hospitals across the United States between 2014 and 2015 [12]. Because it is multicenter, eICU is commonly used for external validation of clinical prediction models.

Both databases are fully deidentified, and all personally identifiable information has been removed. Accordingly, this study was a secondary analysis of publicly available data and did not require additional ethical approval under applicable regulations.

Both cohorts had the same inclusion criteria: age ≥ 18 years, use of invasive mechanical ventilation, and length of ICU stay ≥ 48 h. The first step in the missing-data screening was to remove candidate variables with more than 30 percent missing values. The missingness by a patient was subsequently calculated on the remaining candidate variables and patients with more than 30 percent missing data was removed. Also, patients with hematologic malignancies and those who died in the hospital were excluded. In case of repeated ICU admissions, only the first ICU admission was included, and other admissions were eliminated.

The clinical measures measured within 24 h prior to ICU admission were taken out as the development cohort. Since it was not possible to get continuous data on the 24 h before ICU admission in the eICU, the initial recorded measurement of each respective variable upon entering the ICU was used to validate externally. Following the previously mentioned missing data exclusion at the variable level and patient level, the rest of the missing values were replaced by random forest imputation separately in development and external validation cohorts.

A total of 1601 patients were included in the development cohort and 1576 in the external validation cohort. The MIMIC-IV cohort was used

sequentially for baseline characteristic analysis, missing-data imputation, feature selection, development of six machine-learning models, 10-fold cross-validation, and SHAP analysis. The eICU cohort was used exclusively for independent external validation of the final model. The patient selection process is shown in Fig. 1.

2.2. Definition of pressure injury and outcome ascertainment

The outcome of interest was the presence or absence of pressure injury-related diagnostic records in the electronic health record during the ICU stay. In MIMIC-IV, pressure injuries were identified using International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes 707.0 $\times$ and 707.2 $\times$ and Tenth Revision, Clinical Modification (ICD-10-CM) code category L89. In eICU, ICD codes and diagnostic text were used. Patients were classified as having documented pressure injury if a pressure injury-related diagnosis was recorded during the hospitalization corresponding to the index ICU stay. Outcomes were ascertained from structured diagnostic information without additional clinical adjudication.

2.3. Study variables

The demographic characteristics (age, sex, race/ethnicity, and marital status), the clinical characteristics (height, weight, smoking history, alcohol consumption history, comorbidities, first hospital admission, sepsis, admission type, and use of sedatives, neuromuscular blocking agents, vasopressors, and glucocorticoids) and the laboratory variables (mean values of white blood cell count, platelet count, hematocrit, mean corpuscular volume, blood glucose, blood urea nitrogen, serum potassium, and anion gap) were taken as variables in MIMIC-IV during the 24 h preceding the ICU admission. Candidate variables that had more than 30 percent of missing values were removed first. Missingness at the patient level was then computed based on the remaining candidate variables and patients with more than 30 percent missingness were dropped. Missing values were then imputed with random forest imputation. The subsequent MIMIC-IV development cohort was subsequently applied to select features and develop models.

2.4. Feature selection

Feature selection was done independently with the help of the Boruta algorithm (Fig. 2) and LASSO regression (Fig. 3), after random forest imputation. Intersection of the features chosen by the two methods formed the first candidate predictor set. Preliminary classification model was created thereafter, and SHAP analysis was employed to estimate the contribution of each variable into the model output. Manual trimming was then constrained to clinical interpretability, database transportability, and agreement with previous evidence. Elective admission, glucocorticoid use, and sedative use were eliminated as these variables represent treatment or admission procedures which can differ among hospitals and could not be more appropriate to be included in a transportable EHR-based model upon the examination of their SHAP contribution and clinical significance. White blood cell count and blood glucose were included due to the fact that prior investigations have shown relationships with pressure injury [4,13,14] and both measures are readily available with similar definitions in MIMIC-IV and eICU. The final classification model consisted of six features, i.e., sex, white blood cell count, blood glucose, hematocrit, blood urea nitrogen, and anion gap.

2.5. Model development and validation

Using the selected features, six machine-learning models were initially compared in the MIMIC-IV development cohort: extreme gradient boosting (XGBoost), adaptive boosting (AdaBoost), Light Gradient Boosting Machine (LightGBM), random forest (RF), gradient-

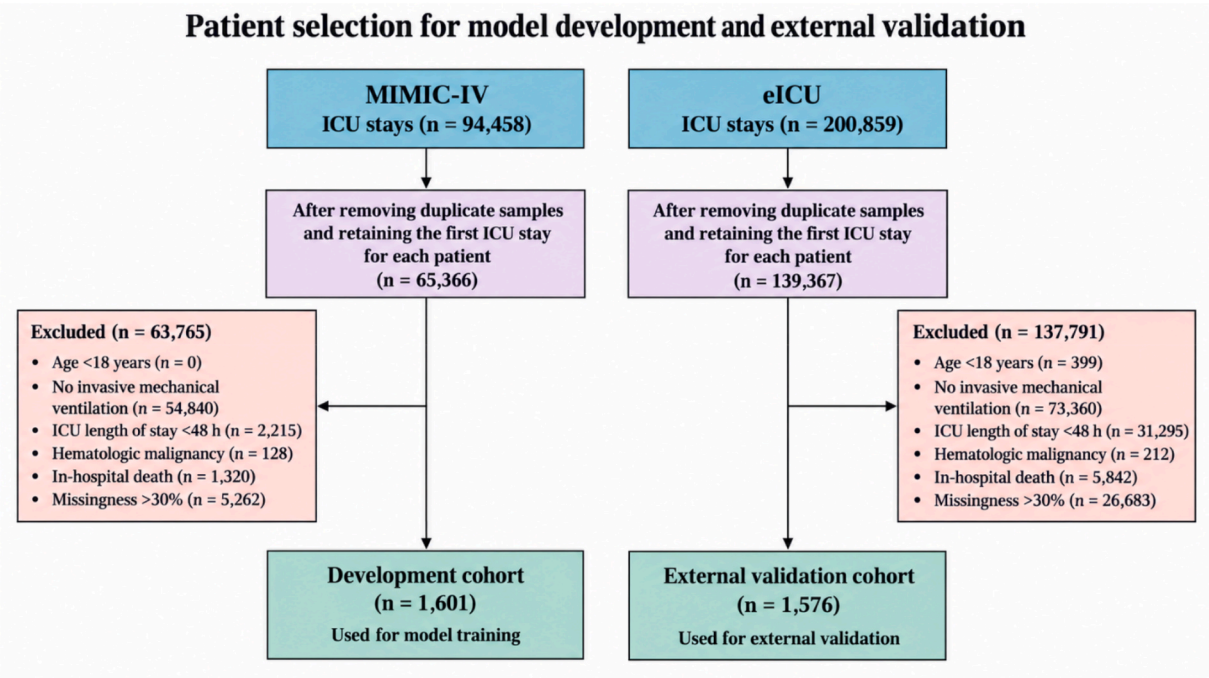


Fig. 1. Patient selection process.

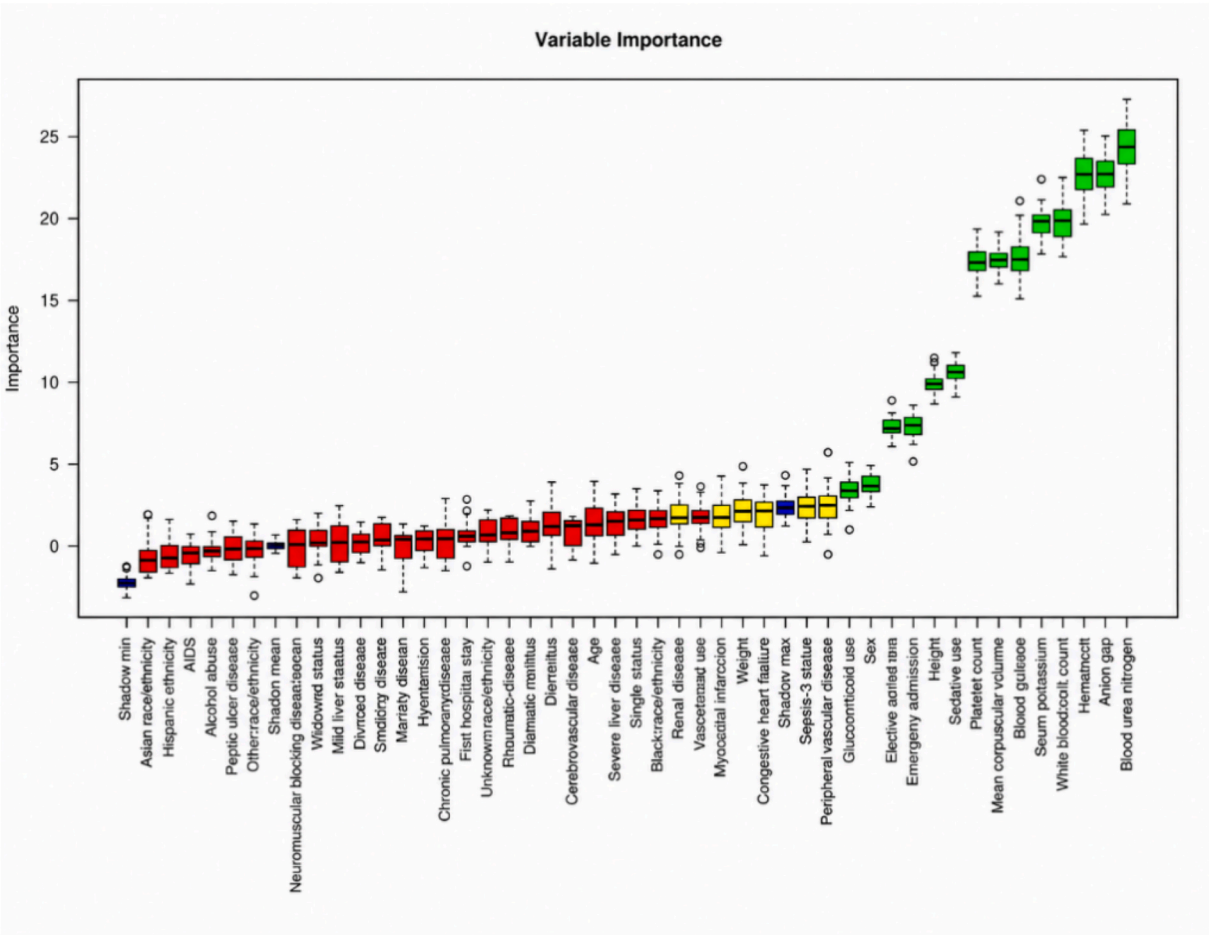


Fig. 2. Feature selection using the Boruta algorithm.

boosting decision tree (GBDT), and decision tree (DT). The hyperparameter settings used in the initial comparison are provided in

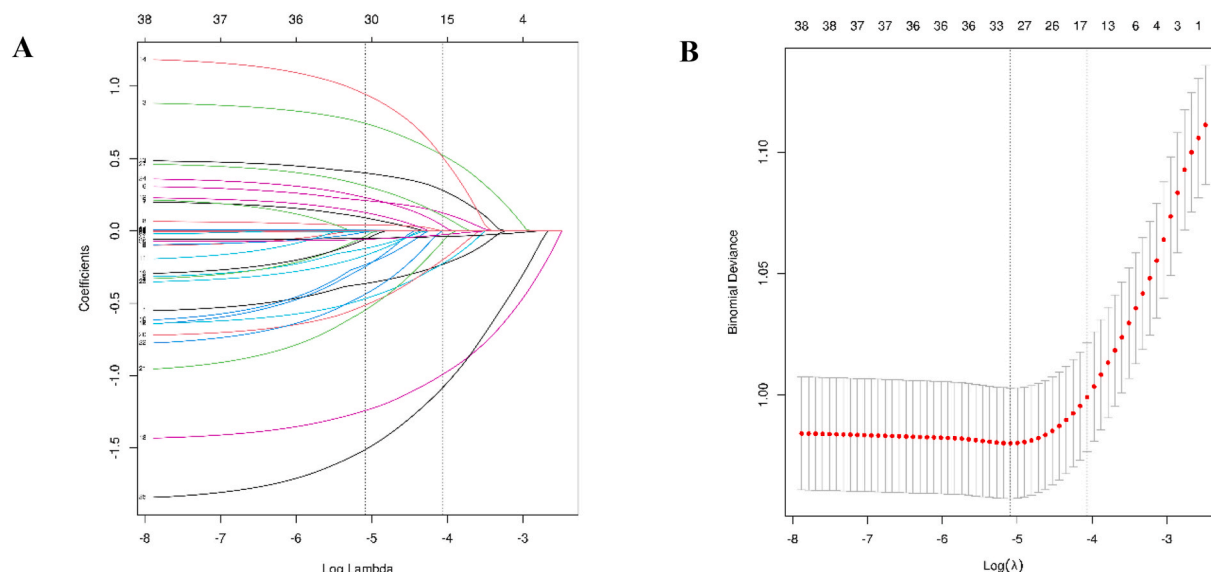


Fig. 3. Feature selection using LASSO regression.

Supplementary Table S1. The area under the receiver operating characteristic curve (AUC) was the primary performance metric. Additional performance assessments included sensitivity, specificity, accuracy, F1 score, Cohen's kappa, calibration, precision–recall analysis, and decision curve analysis.

In the first six-model comparison, the 10-fold cross-validation was used to measure model performance. Following the choice of XGBoost as the highest-performing candidate model, the specific optimization analysis of the XGBoost was conducted. This was done by randomly holding out 15.05% of the MIMIC-IV population ($n = 241$) as an internal test sample and performing 10-fold cross-validation with GridSearchCV in the remaining samples. ROC-AUC was the GridSearchCV optimization metric and the random seed was 42. The optimal values of XGBoost parameters were `colsample_bytree = 1`, `learning_rate = 0.1`, `max_depth = 4`, `min_child_weight = 2`, `n_estimators = 20`, `reg_lambda = 1`, and `subsample = 1`. Fixed probability threshold of 0.304 which was determined in the process of developing the model was used to compute sensitivity, specificity, accuracy, F1 score, and confusion matrices and kept to validate externally. Subsequently, the optimized XGBoost was applied to the eICU cohort to be validated independently externally. A calibration plot and the Brier score were used to determine external calibration. Cross-validation was preceded by random forest imputation and feature selection that was done once and was not repeated in every fold so that the results of the internal cross-validation can be viewed as conditional on the pre-processed and pre-selected predictor set rather than as estimates of a fully nested processing pipeline.

2.6. Statistical analysis

Summary and comparison of baseline characteristics between groups were made across the MIMIC-IV development cohort. Each of the statistical analyses was carried out with R version 4.2.3 and Python version 3.11.4. Continuous variables with normal distribution (mean plus/minus standard deviation) were compared by independent-samples *t*-test. Continuous variables that were non-normally distributed (median [interquartile range]) were compared by Mann–Whitney *U* test. Counts and percentages were used to summarize categorical variables and Pearson chi-square test or Fisher exact test was used based on the situation. Absolute standardized mean differences (SMDs) were used to compare the distributions of the six final predictors between MIMIC-IV and eICU to describe dataset shift where higher numbers represent

greater between-cohort imbalance. The 95% confidence intervals of sensitivity, specificity, accuracy, positive predictive value, and negative predictive value were determined through use of Wilson score intervals; the 95% confidence interval of the F1 score was estimated through the application of multinomial bootstrap resampling of the 2×2 confusion matrix (200,000 resamples). Overall agreement between predicted probabilities and observed outcomes was assessed using a calibration plot and the Brier score and lower scores indicated better overall agreement between predicted probabilities and observed outcomes.

3. Results

3.1. Baseline characteristics

After applying the inclusion and exclusion criteria, 1601 patients were included in the development cohort. Patients were grouped according to the presence or absence of pressure injury-related diagnoses documented in the electronic health record during the ICU stay. Baseline characteristics are presented in Table 1.

Several baseline characteristics differed between groups. Compared with patients without documented pressure injury, those with documented pressure injury were older and had lower median hematocrit and higher median blood urea nitrogen levels. Other categorical and laboratory differences are shown in Table 1.

3.2. Comparison of the performance of multiple machine-learning algorithms

Comparison of six different machine learning algorithms was done on the MIMIC-IV development cohort regarding prediction of the presence of documented pressure injury during hospitalization in ICU patients on invasive mechanical ventilation. At this first model-comparison step, XGBoost had the best AUC of 0.866 (95% CI, 0.787–0.944), followed by AdaBoost (AUC = 0.861), RF (AUC = 0.853), DT (AUC = 0.839), GBDT (AUC = 0.828), and LightGBM (AUC = 0.724) (Fig. 4A and Table 2).

Fig. 4B displays decision curve analysis findings. XGBoost has a large net benefit at most of the threshold probabilities. The calibration curve of XGBoost was nearest to the diagonal reference line as shown in Fig. 4C. Precision-recall curves are depicted in Fig. 4D. XGBoost had the maximum average precision (AP = 0.815); RF and AdaBoost also demonstrated good precision-recall performance (AP = 0.811 and 0.794

Table 1

Baseline characteristics of the development cohort.

Variable	Overall (n = 1601)	No documented pressure injury (n = 1211)	Documented pressure injury (n = 390)	Test statistic	p value
Sex: Male, n (%)	906(56.590)	657(54.253)	249(63.846)	11.052	<0.001
Sex: Female, n (%)	695(43.410)	554(45.747)	141(36.154)		
First hospital stay: No, n (%)	806(50.344)	600(49.546)	206(52.821)	1.265	0.261
First hospital stay: Yes, n (%)	795(49.656)	611(50.454)	184(47.179)		
Sepsis-3 status: No, n (%)	254(15.865)	225(18.580)	29(7.436)	27.445	<0.001
Sepsis-3 status: Yes, n (%)	1347(84.135)	986(81.420)	361(92.564)		
Smoking history: No, n (%)	1426(89.069)	1071(88.439)	355(91.026)	2.027	0.155
Smoking history: Yes, n (%)	175(10.931)	140(11.561)	35(8.974)		
Alcohol abuse: No, n (%)	1582(98.813)	1198(98.927)	384(98.462)	0.544	0.461
Alcohol abuse: Yes, n (%)	19(1.187)	13(1.073)	6(1.538)		
Hypertension: No, n (%)	913(57.027)	674(55.656)	239(61.282)	3.810	0.051
Hypertension: Yes, n (%)	688(42.973)	537(44.344)	151(38.718)		
Diabetes mellitus: No, n (%)	1067(66.646)	839(69.282)	228(58.462)	15.536	<0.001
Diabetes mellitus: Yes, n (%)	534(33.354)	372(30.718)	162(41.538)		
Myocardial infarction: No, n (%)	1378(86.071)	1054(87.036)	324(83.077)	3.856	0.050
Myocardial infarction: Yes, n (%)	223(13.929)	157(12.964)	66(16.923)		
Congestive heart failure: No, n (%)	1094(68.332)	850(70.190)	244(62.564)	7.928	0.005
Congestive heart failure: Yes, n (%)	507(31.668)	361(29.810)	146(37.436)		
Mild liver disease: No, n (%)	1330(83.073)	1011(83.485)	319(81.795)	0.599	0.439
Mild liver disease: Yes, n (%)	271(16.927)	200(16.515)	71(18.205)		
Severe liver disease: No, n (%)	1481(92.505)	1121(92.568)	360(92.308)	0.029	0.865
Severe liver disease: Yes, n (%)	120(7.495)	90(7.432)	30(7.692)		
Renal disease: No, n (%)	1197(74.766)	923(76.218)	274(70.256)	5.557	0.018
Renal disease: Yes, n (%)	404(25.234)	288(23.782)	116(29.744)		
Chronic pulmonary disease: No, n (%)	1075(67.146)	818(67.547)	257(65.897)	0.364	0.546
Chronic pulmonary disease: Yes, n (%)	526(32.854)	393(32.453)	133(34.103)		
Cerebrovascular disease: No, n (%)	1351(84.385)	1032(85.219)	319(81.795)	2.625	0.105
Cerebrovascular disease: Yes, n (%)	250(15.615)	179(14.781)	71(18.205)		
Peripheral vascular disease: No, n (%)	1369(85.509)	1056(87.201)	313(80.256)	11.480	<0.001
Peripheral vascular disease: Yes, n (%)	232(14.491)	155(12.799)	77(19.744)		
Dementia: No, n (%)	1566(97.814)	1192(98.431)	374(95.897)	8.856	0.003
Dementia: Yes, n (%)	35(2.186)	19(1.569)	16(4.103)		
Peptic ulcer disease: No, n (%)	1543(96.377)	1170(96.614)	373(95.641)	0.800	0.371
Peptic ulcer disease: Yes, n (%)	58(3.623)	41(3.386)	17(4.359)		
AIDS: No, n (%)	1575(98.376)	1189(98.183)	386(98.974)	1.155	0.282
AIDS: Yes, n (%)	26(1.624)	22(1.817)	4(1.026)		
Rheumatic disease: No, n (%)	1527(95.378)	1154(95.293)	373(95.641)	0.081	0.776
Rheumatic disease: Yes, n (%)	74(4.622)	57(4.707)	17(4.359)		
Glucocorticoid use: No, n (%)	1465(91.505)	1093(90.256)	372(95.385)	9.982	0.002
Glucocorticoid use: Yes, n (%)	136(8.495)	118(9.744)	18(4.615)		
Sedative use: No, n (%)	1246(77.826)	887(73.245)	359(92.051)	60.457	<0.001
Sedative use: Yes, n (%)	355(22.174)	324(26.755)	31(7.949)		
Neuromuscular blocking agent use: No, n (%)	1588(99.188)	1199(99.009)	389(99.744)	1.976	0.160
Neuromuscular blocking agent use: Yes, n (%)	13(0.812)	12(0.991)	1(0.256)		
Vasopressor use: No, n (%)	1532(95.690)	1152(95.128)	380(97.436)	3.810	0.051
Vasopressor use: Yes, n (%)	69(4.310)	59(4.872)	10(2.564)		
Married status: No, n (%)	806(50.344)	596(49.216)	210(53.846)	2.530	0.112
Married status: Yes, n (%)	795(49.656)	615(50.784)	180(46.154)		
Single status: No, n (%)	1107(69.144)	858(70.851)	249(63.846)	6.784	0.009
Single status: Yes, n (%)	494(30.856)	353(29.149)	141(36.154)		
Divorced status: No, n (%)	1458(91.068)	1096(90.504)	362(92.821)	1.947	0.163
Divorced status: Yes, n (%)	143(8.932)	115(9.496)	28(7.179)		
Widowed status: No, n (%)	1432(89.444)	1083(89.430)	349(89.487)	0.001	0.975
Widowed status: Yes, n (%)	169(10.556)	128(10.570)	41(10.513)		
Black race/ethnicity: No, n (%)	1417(88.507)	1059(87.448)	358(91.795)	5.479	0.019
Black race/ethnicity: Yes, n (%)	184(11.493)	152(12.552)	32(8.205)		
Asian race/ethnicity: No, n (%)	1549(96.752)	1167(96.367)	382(97.949)	2.350	0.125
Asian race/ethnicity: Yes, n (%)	52(3.248)	44(3.633)	8(2.051)		
Hispanic ethnicity: No, n (%)	1549(96.752)	1168(96.449)	381(97.692)	1.451	0.228
Hispanic ethnicity: Yes, n (%)	52(3.248)	43(3.551)	9(2.308)		
Other race/ethnicity: No, n (%)	1387(86.633)	1054(87.036)	333(85.385)	0.694	0.405
Other race/ethnicity: Yes, n (%)	214(13.367)	157(12.964)	57(14.615)		
Elective admission: No, n (%)	1409(88.007)	1030(85.054)	379(97.179)	41.097	<0.001
Elective admission: Yes, n (%)	192(11.993)	181(14.946)	11(2.821)		
Emergency admission: No, n (%)	192(11.993)	181(14.946)	11(2.821)	41.097	<0.001
Emergency admission: Yes, n (%)	1409(88.007)	1030(85.054)	379(97.179)		
Height, cm, median (IQR)	170.000 (163.000–178.000)	168.000 (163.000–178.000)	170.000 (160.000–178.000)	–1.306	0.191
Weight, kg, median (IQR)	80.000 (66.500–96.400)	80.000 (66.800–95.600)	80.000 (65.000–97.000)	0.582	0.561
White blood cell count, $\times 10^9/L$, median (IQR)	10.500 (7.400–14.500)	10.550 (7.400–14.500)	10.000 (7.650–14.300)	0.001	0.999

(continued on next page)

Table 1 (continued)

Variable	Overall (n = 1601)	No documented pressure injury (n = 1211)	Documented pressure injury (n = 390)	Test statistic	p value
Platelet count, $\times 10^9/L$, median (IQR)	221.000 (153.000–302.000)	221.000 (151.000–301.000)	223.000 (158.000–326.000)	−1.019	0.308
Hematocrit, %, median (IQR)	33.100 (28.525–38.100)	33.200 (28.600–38.300)	31.600 (27.700–36.200)	2.562	0.010
Mean corpuscular volume, fL, median (IQR)	91.500 (87.000–96.000)	91.000 (87.000–95.500)	92.000 (88.000–97.000)	−1.353	0.176
Blood glucose, mg/dL, median (IQR)	127.500 (104.000–161.000)	128.000 (105.000–161.000)	120.000 (99.000–161.000)	1.172	0.241
Blood urea nitrogen, mg/dL, median (IQR)	21.000 (14.000–33.000)	20.000 (13.000–31.500)	31.500 (20.000–50.000)	−6.390	<0.001
Serum potassium, mmol/L, median (IQR)	4.100 (3.700–4.500)	4.100 (3.700–4.500)	4.100 (3.800–4.600)	−1.135	0.256
Anion gap, mmol/L, median (IQR)	15.000 (12.000–17.000)	15.000 (12.500–17.000)	14.000 (12.000–17.000)	0.583	0.559
Age, years, median (IQR)	65.330 (55.450–76.320)	64.610 (55.340–75.400)	67.050 (56.020–78.210)	−2.447	0.014

Note: Baseline characteristics were summarized using observed values before imputation.

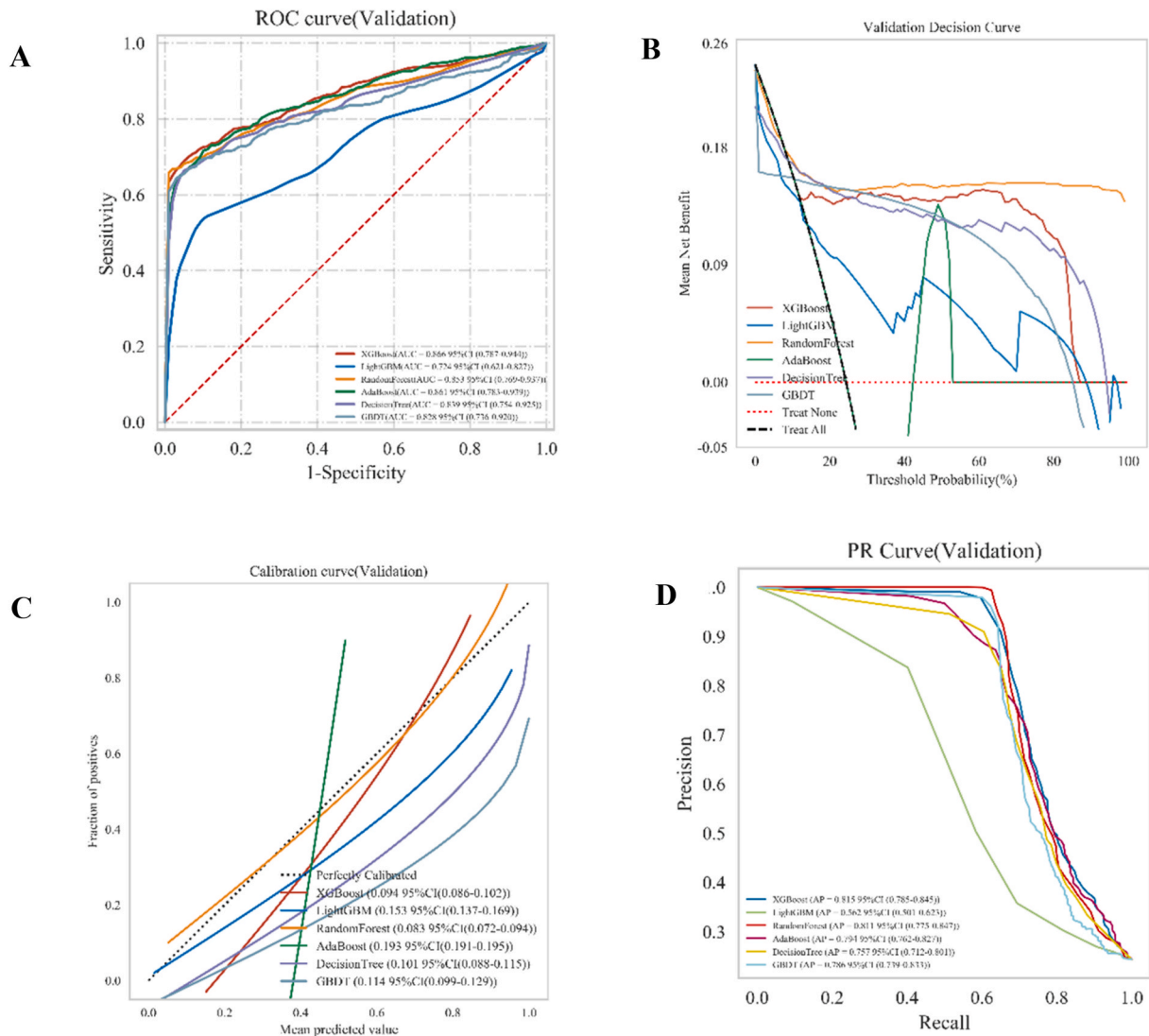


Fig. 4. Comparative performance of the six machine-learning models during the initial model-comparison stage: (A) receiver operating characteristic curves; (B) decision curve analysis; (C) calibration curves; and (D) precision–recall curves.

respectively). The detailed performance measures are given in Table 2.

The comparison of models based on model discrimination was also done through the use of the DeLong test (Table 3). There were no statistically significant pairwise comparisons between XGBoost (all $p > 0.05$), but this method had the greatest point-estimate AUC and performed well in all three aspects: discrimination, calibration and

decision curve analyses. XGBoost was thus chosen as an object of further optimization.

Upon selecting the XGBoost model, a specific XGBoost optimization analysis followed. This analysis involved reserving 15.05 percent of the MIMIC-IV cohort ($n = 241$) as an internal test set and the rest of the samples were subjected to 10-fold cross-validation as part of

Table 2

Detailed performance metrics of the machine-learning models.

Classification model	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	F1 score (95% CI)	Cohen's κ (95% CI)
XGBoost	0.866 (0.787–0.944)	0.738 (0.690–0.787)	0.889 (0.857–0.922)	0.711 (0.672–0.751)	0.613 (0.557–0.669)
LightGBM	0.724 (0.621–0.827)	0.551 (0.497–0.606)	0.905 (0.822–0.988)	0.610 (0.555–0.664)	0.499 (0.411–0.587)
Random forest	0.853 (0.769–0.937)	0.636 (0.579–0.692)	0.995 (0.991–0.999)	0.767 (0.722–0.811)	0.713 (0.662–0.764)
AdaBoost	0.861 (0.783–0.939)	0.662 (0.607–0.716)	0.963 (0.946–0.979)	0.743 (0.697–0.789)	0.675 (0.618–0.732)
Decision tree	0.839 (0.754–0.925)	0.726 (0.674–0.778)	0.856 (0.827–0.884)	0.668 (0.632–0.704)	0.549 (0.501–0.598)
GBDT	0.828 (0.736–0.920)	0.631 (0.572–0.690)	0.989 (0.980–0.999)	0.755 (0.704–0.806)	0.697 (0.638–0.757)

Table 3

Pairwise p values from DeLong's test for comparison of model AUCs.

Model	XGBoost	LightGBM	Random forest	AdaBoost	Decision tree	GBDT
XGBoost	NA	0.054	0.526	0.505	0.493	0.268
LightGBM	0.054	NA	0.119	0.060	0.127	0.309
Random forest	0.526	0.119	NA	0.569	0.521	0.414
AdaBoost	0.505	0.060	0.569	NA	0.486	0.425
Decision tree	0.493	0.127	0.521	0.486	NA	0.509
GBDT	0.268	0.309	0.414	0.425	0.509	NA

GridSearchCV. The optimized XGBoost model recorded 10-fold cross-validated AUC of 0.853 (95% CI, 0.765–0.941) (Fig. 5). Hence, the AUC of 0.866 obtained in the first six-model comparison and the AUC of 0.853 obtained in the later optimized XGBoost analysis are two different stages of model development instead of two different summaries of the same cross-validation process. A Kolmogorov-Smirnov (KS) plot of the optimized XGBoost model is shown in Fig. 6 with a KS statistic of 0.741.

3.3. SHAP analysis

The SHAP analysis was done in order to explain the way every feature affected the optimized XGBoost model output. As shown in the SHAP beeswarm plot (Fig. 7), low anion gap values and high blood urea nitrogen values caused models to shift their output towards the positive class of documented pressure injury. Hematocrit also showed a substantial contribution to model output, although its SHAP pattern did not

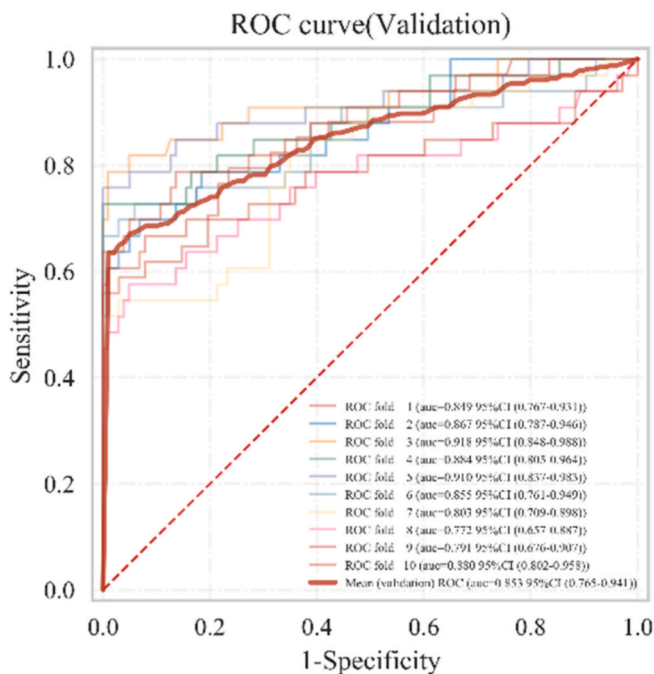


Fig. 5. Fold-specific receiver operating characteristic curves for the optimized XGBoost model across the 10 cross-validation folds; the cross-validated AUC was 0.853 (95% CI, 0.765–0.941).

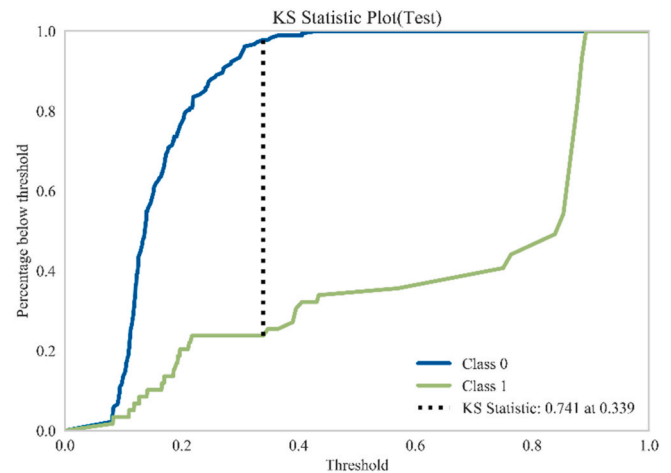


Fig. 6. Kolmogorov-Smirnov (KS) plot for the optimized XGBoost model in the MIMIC-IV development cohort.

demonstrate a simple monotonic relationship with documented pressure injury-positive classification. Conversely, the number of white blood cells and blood glucose did not display straightforward monotone correlations with the model output. The SHAP value of sex was clustered around zero, which is indicative of a relatively small impact on classification.

Fig. 8 contains the ranking of features based on SHAP. The absolute SHAP was sorted in the following order (highest to lowest): anion gap, blood urea nitrogen, hematocrit, white blood cell count, blood glucose and sex. The highest contributions to the model output were anion gap and blood urea nitrogen, whereas blood glucose and sex contributed much less.

3.4. External validation

Table 4 provides the summary of the distributions of the six final predictors in the MIMIC-IV development cohort and in the eICU external validation cohort. The biggest differences between cohorts were found in anion gap (absolute SMD = 0.533) and hematocrit (absolute SMD = 0.441), and then in blood glucose (absolute SMD = 0.275). Blood urea nitrogen (absolute SMD = 0.197) and sex (absolute SMD = 0.194) had smaller differences, but white blood cell count was comparatively close between cohorts (absolute SMD = 0.065).

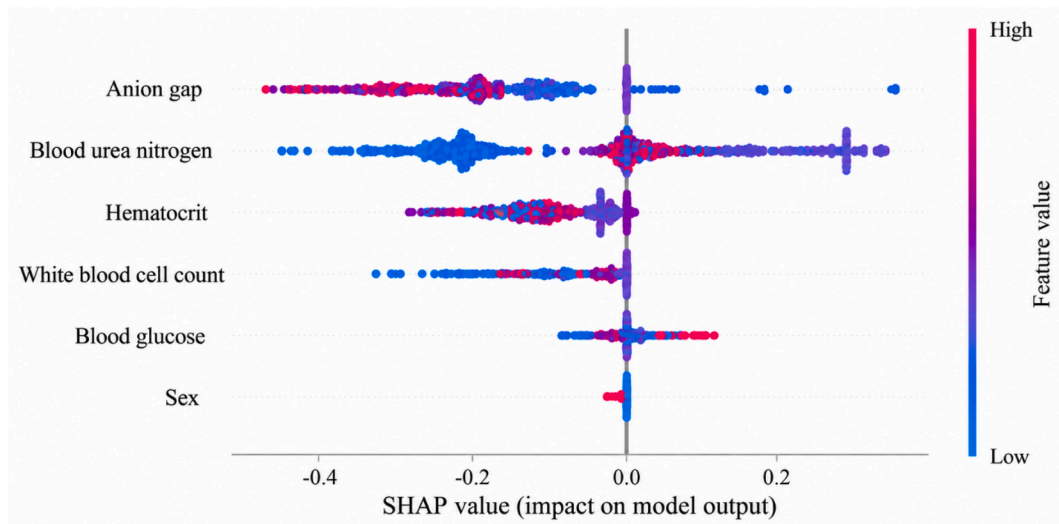


Fig. 7. SHAP beeswarm plot.

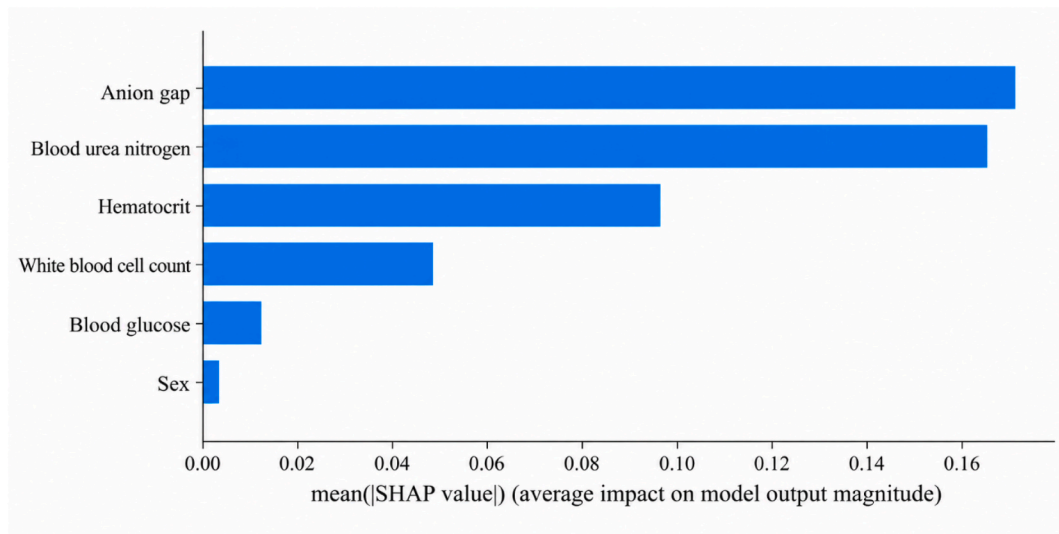


Fig. 8. SHAP-based feature importance ranking.

Table 4

Distribution of the six final predictors in the MIMIC-IV development and eICU external validation cohorts.

Predictor	MIMIC-IV development cohort (n = 1601)	eICU external validation cohort (n = 1576)	Absolute SMD
Sex, n (%)	695 (43.4%)	836 (53.0%)	0.194
White blood cell count, $\times 10^9/L$	10.50 (7.40–14.49)	11.10 (8.20–15.00)	0.065
Blood glucose, mg/dL	127.42 (104.00–161.00)	145.00 (112.00–201.00)	0.275
Hematocrit, %	33.10 (28.53–38.10)	36.10 (31.60–41.00)	0.441
Blood urea nitrogen, mg/dL	21.00 (14.00–33.00)	28.00 (19.00–44.00)	0.197
Anion gap, mmol/L	15.00 (12.00–17.00)	12.00 (9.00–15.00)	0.533

Data are presented as n (%) for sex and median (IQR) for continuous variables. Continuous-variable summaries and SMDs use observed values before imputation. SMD, standardized mean difference.

The final optimized XGBoost model was then applied to the

independent eICU cohort. Among 1576 patients, 221 (14.0%) had documented pressure injury and 1355 (86.0%) did not. The model achieved an AUC of 0.703 (95% CI, 0.669–0.737) (Fig. 9). At the fixed probability threshold of 0.304, the confusion matrix included 1146 true negatives, 209 false positives, 153 false negatives, and 68 true positives (Fig. 10). Sensitivity was 0.308 (95% CI, 0.251–0.371), specificity was 0.846 (95% CI, 0.826–0.864), accuracy was 0.770 (95% CI, 0.749–0.790), positive predictive value was 0.245 (95% CI, 0.199–0.299), negative predictive value was 0.882 (95% CI, 0.864–0.899), and the F1 score was 0.273 (95% CI, 0.221–0.324). External calibration yielded a Brier score of 0.119 (95% CI, 0.109–0.129). The calibration curve showed deviation from perfect calibration, particularly at higher predicted probabilities (Fig. 11).

4. Discussion

In this study, MIMIC-IV version 3.1 was used as the development cohort and eICU version 2.0 as the independent external validation cohort. After missing-data handling and feature selection, six candidate machine-learning algorithms were compared. XGBoost showed the best overall performance in the initial comparison and was subsequently

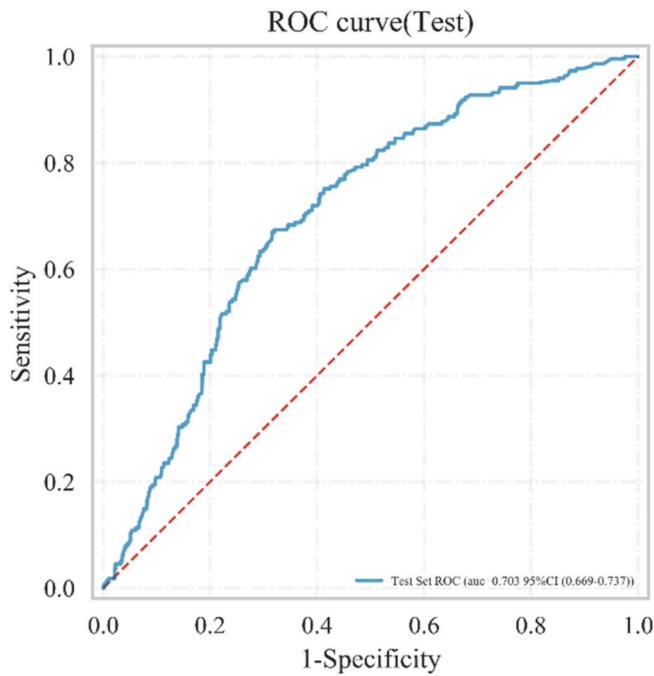


Fig. 9. Receiver operating characteristic curve for the independent eICU external validation cohort.

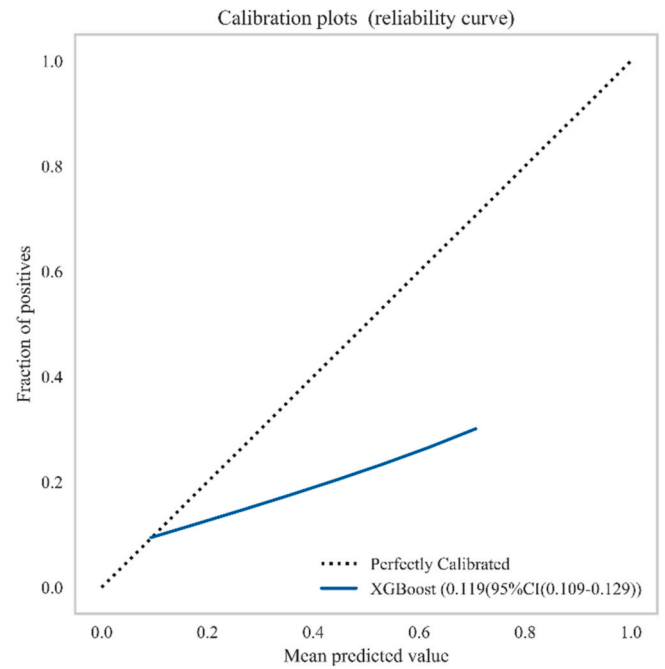


Fig. 11. Calibration plot of the final XGBoost model in the independent eICU external validation cohort. The diagonal dotted line represents perfect calibration. The Brier score was 0.119 (95% CI, 0.109–0.129).

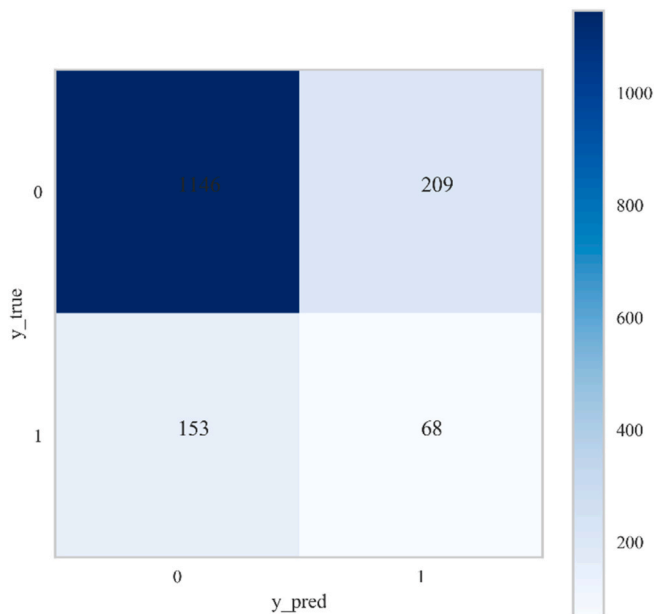


Fig. 10. Confusion matrix for the final XGBoost model in the eICU external validation cohort at the fixed probability threshold of 0.304.

optimized using GridSearchCV. The final optimized XGBoost model achieved a 10-fold cross-validated AUC of 0.853 in the MIMIC-IV cohort and an AUC of 0.703 in the eICU cohort. SHAP analysis indicated that anion gap and blood urea nitrogen contributed most strongly to model output. External calibration yielded a Brier score of 0.119.

Patients receiving invasive mechanical ventilation are at high risk of pressure injury because of severe illness, restricted mobility, and frequent exposure to sedative therapy [5]. Lima-Serrano et al. [15], identified impaired activity or mobility, compromised tissue perfusion, and vasopressor use as relatively consistent risk factors among critically ill patients. An updated 2025 review [13] further confirmed the

importance of restricted mobility, impaired perfusion, vasopressor use, and prolonged ICU length of stay. However, invasive mechanical ventilation should not be viewed simply as a direct cause of pressure injury. Rather, ICU patients requiring invasive mechanical ventilation generally have greater illness severity, prolonged immobility, greater exposure to sedation and analgesia, poorer tissue perfusion, and more frequent contact with medical devices. Lee et al. [16] reported that longer exposure to neuromuscular blocking agents and opioids was associated with increased pressure injury risk, possibly because these treatments contribute to immobility. A systematic review of pressure injuries in critical care patients [17] also identified prolonged device use, vasoconstrictor administration, and prone-position ventilation as potential risk factors.

In addition, several factors that were not retained in the final six-feature model, including sedative use, vasopressor use, glucocorticoid use, and certain comorbidities, still showed some degree of importance in the Boruta analysis and were broadly consistent with previous findings. This may be explained by the different objectives of the Boruta and LASSO methods. Boruta is designed to identify as many relevant variables as possible, whereas LASSO places greater emphasis on model parsimony and retains the most informative predictors. Therefore, exclusion of a factor from the final model does not necessarily indicate that it is unrelated to pressure injury. Moreover, medication exposure was coded only as a binary yes/no variable in this study, without accounting for dosage, treatment duration, drug class, or clinical indication. This simplified definition may have underestimated the contribution of medication-related factors to the model.

Machine-learning methods can capture complex higher-order interactions and nonlinear relationships among clinical variables and have been widely applied to pressure injury risk prediction. A 2023 systematic review and meta-analysis [6] found relatively high pooled discrimination for machine-learning pressure injury models. However, the included studies were generally limited by single-center designs, small sample sizes, incomplete model reporting, and inadequate external validation. Another systematic review [18] found substantial heterogeneity in feature selection, missing-data handling, and model evaluation methods.

A 2025 MIMIC-IV study [10] investigated pressure injury prediction among mechanically ventilated ICU patients. The study included 29,448 patients and developed an XGBoost model interpreted using SHAP. The model achieved an AUC of 0.739 in internal testing. The most influential variables included sepsis, age, platelet count, ICU length of stay, oxygenation index, hemoglobin concentration, admission type, renal disease, albumin, and race. However, the model was not validated in an independent external database. Compared with that study, the present study compared multiple machine-learning algorithms, evaluated the optimal model using 10-fold cross-validation, and performed independent external validation using the multicenter eICU dataset.

The last optimized XGBoost model obtained a cross-validated AUC of 0.853 in the MIMIC-IV development group and an AUC of 0.703 in the eICU external validation group. The decrease in discrimination can be due to the disparity in patient characteristics between the cohorts, prevalence of outcomes, hospital environment, clinical practice behavior, and data collection processes. The eICU database has records of various hospitals, whereas MIMIC-IV is based on a single centre. Such differences render external validation more difficult, but also offer a more stringent evaluation of model transportability.

One other reason that could have led to the poor external validation performance was the lack of consistency in the predictor measurement windows. Predictors were measured as mean values collected over the 24 h preceding the ICU admission in MIMIC-IV, but the first available measurements following ICU admission in eICU as pre-ICU data could not be used. The comparison of the six final predictors demonstrated measurable distributional differences between the two cohorts. The highest absolute SMDs were seen on anion gap (0.533) and hematocrit (0.441), followed by blood glucose (absolute SMD = 0.275); blood urea nitrogen and sex had smaller differences (0.197 and 0.194, respectively) and white blood cell count has relatively small difference (0.065). Moreover, the recorded number of pressure injuries in eICU was less than in MIMIC-IV (14.0% vs 24.4%). Such disparities in measurement windows, predictor distributions, and outcome prevalence can also in part account for why the level of external discrimination decreased.

The external operating characteristics are also providing extra information on the clinical interpretation. The sensitivity and specificity values at a fixed threshold of 0.304 were 0.308 and 0.846 respectively, which means that a significant number of patients with documented pressure injuries were not positively labeled. The given operating point cannot thus be used as an individual screening device or rule-out tool. Also, even though the Brier score was 0.119, the external calibration curve was off the ideal diagonal line at higher predicted probabilities, which means that there was calibration drift among the cohorts. Whenever the model is subjected to prospective evaluation, threshold selection must correspond to the intended clinical purpose and local recalibration ought to be taken into consideration.

Anion gap was the most influential feature in the SHAP analysis. The anion gap is calculated as follows: sodium – (chloride + bicarbonate). SHAP analysis showed that a lower anion gap was associated with a higher probability of documented pressure injury-positive classification. This finding appears inconsistent with the conventional understanding that high anion gap metabolic acidosis generally reflects greater illness severity. One plausible explanation is that hypoalbuminemia lowers the measured anion gap because albumin is the most abundant unmeasured anion in plasma; therefore, a reduction in albumin concentration can decrease the calculated anion gap [19,20]. Hypoalbuminemia is also associated with illness severity, poor nutritional status, and systemic inflammation. Accordingly, the anion gap should not be interpreted simply as an absolute indicator of metabolic acidosis. Protein levels and other metabolic factors, as well as their interactions with other model features, should also be considered.

Blood urea nitrogen also made a substantial contribution to model output. In the SHAP analysis, higher blood urea nitrogen levels were associated with an increased probability of documented pressure injury-positive classification. Elevated blood urea nitrogen is commonly

observed in renal dysfunction, impaired tissue perfusion, and hypercatabolic states, all of which may reflect poor nutritional status, compromised microcirculatory perfusion, and reduced tissue-healing capacity. A previous XGBoost-based study of mechanically ventilated patients [10] also identified renal disease as an important feature and suggested that renal impairment may increase the vulnerability of skin and soft tissues through metabolic abnormalities, fluid and electrolyte disturbances, and poor nutritional status.

Hematocrit represents the proportion of circulating blood volume occupied by red blood cells. SHAP analysis indicated that lower hematocrit levels were associated with documented pressure injury-positive classification. A low hematocrit may reflect anemia, blood loss, or hemodilution, all of which may impair tissue oxygen delivery and thereby increase vulnerability of pressure-exposed tissues [21].

White blood cell count contributed to model classification, although its SHAP pattern was non-monotonic. This finding is broadly consistent with Jiang et al. [14], who identified elevated white blood cell count as an independent risk factor. Recent studies have suggested that higher white blood cell count may be associated with greater pressure injury risk. However, most available evidence comes from single-center observational studies, and further research is required to determine whether white blood cell count is an effective independent predictor. Sex also provided some stratification information, although its overall contribution to classification was limited.

The six variables included in the final model are all available electronically through health records and can be automatically extracted and computed. Theoretically, the model might be incorporated within the ICU information systems to offer additional data on evaluation of the documented pressure injury status of patients undergoing invasive mechanical ventilation. Nevertheless, the external operating parameters have an obvious weakness: the sensitivity was found to be 0.308 at the set probability threshold level of 0.304 and the specificity was 0.846. Therefore, the use of the model cannot be a standalone screening or a rule-out device and should not substitute the Braden Scale or judgement by the nurse. When considered prospectively with a view to implementation in clinical practice, the value of the classification threshold must be chosen based on the purpose, with more focus on sensitivity in case the purpose is early clinical evaluation.

The present study has several strengths. First, it focused specifically on ICU patients receiving invasive mechanical ventilation and compared six machine-learning algorithms. Second, feature selection combined Boruta and LASSO, and SHAP analysis improved model interpretability. Third, the model was externally validated in an independent multicenter eICU cohort. Compared with simply dividing MIMIC-IV into training and testing subsets, validation in eICU provides a more rigorous assessment of transportability across hospital settings. The final model also included only six routinely available variables, supporting potential integration into electronic health record systems for automated classification and stratification.

This study also has several limitations. First, because the study was based on retrospective databases, selection bias and information bias could not be completely avoided. The findings reflect only associations and classification relationships between the included variables and documented pressure injury status and should not be interpreted as evidence of causality. Second, the ICD diagnosis records in MIMIC-IV did not provide the actual time of pressure injury onset, while the diagnosis time offsets in eICU primarily reflected the time at which diagnoses were entered rather than the time at which the lesions first developed. Therefore, pressure injuries present before ICU admission could not be strictly distinguished from those newly occurring during the ICU stay. It was also not possible to establish the precise temporal sequence among invasive mechanical ventilation, predictor measurement, and pressure injury documentation.

Third, predictor measurement differed between the development and external validation cohorts. Mean values recorded within 24 h before ICU admission were used in MIMIC-IV, whereas the first

measurements obtained after ICU admission were used in eICU. This inconsistency was accompanied by measurable predictor distribution shifts and may have adversely affected external validation performance. The external calibration curve also deviated from perfect calibration, particularly at higher predicted probabilities, suggesting that local recalibration may be required before clinical implementation in other ICU settings. In addition, several clinically important factors, including the Braden Scale score, mobility, nutritional status, and the use of pressure-redistributing mattresses, were not included in the model. Random forest imputation could not completely eliminate bias caused by data that were not missing at random. Furthermore, missing-data imputation and feature selection were performed before cross-validation and were not re-estimated within each fold. Consequently, information from validation folds may have indirectly influenced preprocessing and feature selection, potentially leading to optimistic internal performance estimates. A fully nested preprocessing sensitivity analysis was not performed; therefore, the internal validation results should be interpreted with caution. Finally, prospective validation has not yet been conducted, and the model's clinical utility requires further evaluation.

5. Conclusion

Using the MIMIC-IV and eICU databases, we developed and externally validated a machine-learning model for classifying documented pressure injury status during hospitalization among ICU patients receiving invasive mechanical ventilation. After hyperparameter optimization, the final XGBoost model achieved a 10-fold cross-validated AUC of 0.853 in MIMIC-IV and an external AUC of 0.703 in eICU. The principal classification features were anion gap, blood urea nitrogen, hematocrit, white blood cell count, blood glucose, and sex. Although the model may provide supplementary information for comprehensive clinical assessment, its relatively low external sensitivity and imperfect calibration preclude its use as a stand-alone screening tool. Prospective validation and local recalibration are required before routine clinical implementation.

Ethics declaration

This study was a secondary analysis of de-identified data from the MIMIC-IV and eICU Collaborative Research Database and was conducted in accordance with applicable laws, regulations, institutional guidelines, and PhysioNet data-use requirements. The MIMIC-IV research resource was reviewed by the Institutional Review Board of Beth Israel Deaconess Medical Center, which approved the data-sharing initiative and granted a waiver of informed consent. The eICU database was de-identified in accordance with the Safe Harbor provisions of the U.S. Health Insurance Portability and Accountability Act (HIPAA). Because the present study involved only previously collected, de-identified data and no direct interaction with participants, no additional ethics committee approval or individual informed consent was required.

Author contributions

Yang Zhao and Siji Chen: Writing–review & editing, Writing–original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. Chang Pang: Investigation, Data curation. Yao Tang: Investigation, Data curation. Muhua Zeng: Methodology, Data curation. Shuwu Lin: Writing–review & editing, Writing–original draft, Supervision, Methodology, Investigation, Funding acquisition, Data curation. All authors have read and agreed to the published version of the manuscript.

Declarations

During the preparation of this work, the authors used ChatGPT (OpenAI) for language editing and polishing; the authors reviewed and edited the output as needed and take full responsibility for the content of the published article.

Funding

The Shenyang Medical College Master's Science and Technology Innovation Fund (Y20250544).

Conflict of interest

The authors declare that they have no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jtv.2026.101058>.

References

- [1] Li Z, Lin F, Thalib L, Chaboyer W. Global prevalence and incidence of pressure injuries in hospitalised adult patients: a systematic review and meta-analysis. *Int J Nurs Stud* 2020;105:103546.
- [2] Labeau SO, Afonso E, Benbenishty J, Blackwood B, Boulanger C, Brett SJ, Calvino-Gunther S, Chaboyer W, Coyer F, Deschepper M, et al. Prevalence, associated factors and outcomes of pressure injuries in adult intensive care unit patients: the DecubICUs study. *Intensive Care Med* 2021;47(2):160–9.
- [3] Padula WV, Delarmente BA. The national cost of hospital-acquired pressure injuries in the United States. *Int Wound J* 2019;16(3):634–40.
- [4] Cox J, Roche S, Murphy V. Pressure injury risk factors in critical care patients: a descriptive analysis. *Adv Skin Wound Care* 2018;31(7):328–34.
- [5] Manzano F, Navarro MJ, Roldán D, Moral MA, Leyva I, Guerrero C, Sanchez MA, Colmenero M, Fernández-Mondejar E. Pressure ulcer incidence and risk factors in ventilated intensive care patients. *J Crit Care* 2010;25(3):469–76.
- [6] Pei J, Guo X, Tao H, Wei Y, Zhang H, Ma Y, Han L. Machine learning-based prediction models for pressure injury: a systematic review and meta-analysis. *Int Wound J* 2023;20(10):4328–39.
- [7] Nakagami G, Yokota S, Kitamura A, Takahashi T, Morita K, Noguchi H, Ohe K, Sanada H. Supervised machine learning-based prediction for in-hospital pressure injury development using electronic health records: a retrospective observational cohort study in a university hospital in Japan. *Int J Nurs Stud* 2021;119:103932.
- [8] Rajkumar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380(14):1347–58.
- [9] Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. *Proceed IEEE Institute Electric Electron Eng* 2016;104(2):444–66.
- [10] Zheng L, Xue YJ, Yuan ZN, Xing XZ. Explainable SHAP-XGBoost models for pressure injuries among patients requiring mechanical ventilation in the intensive care unit. *Sci Rep* 2025;15(1):9878.
- [11] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023;10(1):1.
- [12] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data* 2018;5:180178.
- [13] Alderden J, Brooks KR, Kennerly SM, Yap TL, Dworak E, Cox J. Risk factors for pressure injuries in critical care patients: an updated systematic review. *Int J Nurs Stud* 2025;169:105127.
- [14] Yifan J, Juan L, Rui T, Li W, Yi S, Lina Y, Rong M, Jianlei X, Jianwen X, Kaihong F, et al. Risk factors and predictive model for pressure injuries in ICU patients with a Braden score ≤ 9 based on initial serum biomarkers at admission. *J Tissue Viability* 2025;34(4):100955.
- [15] Lima Serrano M, González Méndez MI, Carrasco Cebollero FM, Lima Rodríguez JS. Risk factors for pressure ulcer development in Intensive Care Units: a systematic review. *Med Intensiva* 2017;41(6):339–46.
- [16] Lee SY, Oh DK, Hong SB, Lim CM, Huh JW. Neuromuscular blocking agents and opioids are major risk factors for developing pressure injuries in patients in the intensive care unit. *Kor J Intern Med* 2022;37(6):1186–94.
- [17] Alderden J, Rondinelli J, Pepper G, Cummins M, Whitney J. Risk factors for pressure injuries among critical care patients: a systematic review. *Int J Nurs Stud* 2017;71:97–114.
- [18] Zhou Y, Yang X, Ma S, Yuan Y, Yan M. A systematic review of predictive models for hospital-acquired pressure injury using machine learning. *Nurs Open* 2023;10(3):1234–46.

- [19] Figge J, Jabor A, Kazda A, Fencel V. Anion gap and hypoalbuminemia. *Crit Care Med* 1998;26(11):1807–10.
- [20] Haber LA, Dhaliwal G, Lo L, Rizzuto G. Evaluating a low anion gap: a practical approach. *Cleve Clin J Med* 2023;90(10):619–23.
- [21] Amiri FA, Zhang J. Tissue oxygenation around capillaries: effects of hematocrit and arteriole oxygen condition. *Bull Math Biol* 2023;85(6):50.