Development and validation of an ensemble learning risk model for sepsis after abdominal surgery

Xin Shu[1], Yujie Li[1], Yiziting Zhu[1], Zhiyong Yang[1], Xiang Liu[1], Xiaoyan Hu[1], Chunyong Yang[1], Lei Zhao[2], Tao Zhu[3], Yuwen Chen[4], Bin Yi[1]*

[1]Department of Anesthesiology, Southwest Hospital, Third Military Medical University, Chongqing, China
[2]Department of Anesthesiology, Xuan Wu Hospital, Capital Medical University, Beijing, China
[3]Department of Anesthesiology, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[4]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science, Chongqing, China

***Corresponding author:**
Bin Yi
Department of Anesthesiology
Southwest Hospital
Third Military Medical University
Chongqing, China
E-mail: yibin1974@163.com

Submitted: 29 November 2023; **Accepted:** 30 May 2024
Online publication: 6 June 2024

Arch Med Sci
DOI: https://doi.org/10.5114/aoms/189505

Abstract

Introduction: Although their importance has gained attention, the clinical applications of methods for screening patients at high risk of sepsis after abdominal surgery have been restricted. Therefore, we aimed to develop and validate models for screening patients at high risk of sepsis after abdominal surgery based on machine learning with routine variables.

Material and methods: The whole dataset was composed of three representative academic hospitals in China and the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. Routine clinical variables were implemented for model development. The Boruta algorithm was applied for feature selection. Afterwards, ensemble learning and eight other conventional algorithms were used for model fitting and validation based on all features and selected features. The area under the receiver operating characteristic curves (ROC AUC), sensitivity, specificity, F1 score, accuracy, net reclassification index (NRI), integrated discrimination improvement (IDI), decision curve analysis (DCA), and calibration curves were used for model evaluation.

Results: A total of 955 patients undergoing abdominal surgery were finally analyzed (sepsis: 285, non-sepsis: 670). After feature selection, the ensemble learning model constructed by integrating k-Nearest Neighbor (KNN) and Support Vector Machine (SVM) yielded the ROC AUC of 0.892 (0.841–0.944) and accuracy of 85.0% on the test data, and the ROC AUC of 0.782 (0.727–0.838) and accuracy of 68.1% on the validation data, which performed best. Albumin, ASA score, neutrophil-lymphocyte ratio, age, and glucose were the top features associated with postoperative sepsis by KNN and SVM.

Conclusions: We developed a new and potential generalizable model to preoperatively screen patients at high risk of sepsis after abdominal surgery, with the advantages of a representative training cohort and routine variables.

Key words: sepsis, machine learning, postoperative complications, perioperative period, risk assessment.

Introduction

Sepsis, a syndrome of physiologic, pathologic, and biochemical abnormalities induced by infection, is a major public health concern [1].

When sepsis occurs after a surgical procedure or during the postoperative hospital stay, it is typically called postoperative or surgical sepsis [2]. Abdominal surgery is one of the types of surgery most susceptible to postoperative sepsis [2–4], due to the long procedure time, bacterial translocation, and immune deficiency. It is reported that the incidence of severe postoperative sepsis has markedly increased, independently of patient demographics, comorbidities, and surgery type [4]. Recently, Brakenridge *et al.* reported that despite the low in-hospital mortality of postoperative sepsis, it may affect the long-term outcome including developing chronic critical illness and higher 12-month mortality, especially for elderly patients [5]. Moreover, the postoperative sepsis-associated higher risk of readmission, reexamination and longer hospital stay would also bring a high economic burden, especially in developing countries [6]. Therefore, strategies for preoperatively screening high-risk patients and specific perioperative care are badly needed.

In the last few decades, numerous teams have tried to develop tools for screening high risk [7–10], early diagnosis [11, 12], or predicting mortality risk [13, 14] of postoperative sepsis. Though most of them achieved good results, the clinical application to preoperative screening of high-risk patients is still restricted. Most of the studies were for predicting the mortality risk of postoperative sepsis, not for the occurrence, let alone for preoperative screening of high-risk patients. qSOFA and SOFA scores were the most frequently used variables for this area; however, they are more suitable for predicting the mortality risk of sepsis patients rather than the onset. Moreover, they were designed for intensive care unit (ICU) patients, not specifically for surgical patients (e.g., they lack surgical-associated factors). Though studies on early identification of postoperative sepsis tried to apply more clinical variables for prediction, they paid more attention to the intra- and post-operative variables [9–12], or specific markers [8, 15]. With the application of machine learning (ML) algorithms to the medical field, ML showed great potential to accurately predict sepsis onset ahead of time [16]. However, the heterogeneity of datasets, different availability of clinical variables, and unequal robustness of algorithms restrict the generalizability of models. For the models to predict postoperative sepsis onset based on ML algorithms, most of the datasets were from a single center, without another dataset for external validation [8–10]. Meanwhile, despite the satisfactory model performance, the variables used in some existing models are not routine [8, 12] or not specific for patients undergoing abdominal surgery [7, 9, 10]. The generalization and robustness of the aforementioned ML-based models are affected by the single source of data, lack of external validation, and non-routine and non-specific variables, which restrict the application to clinical situations.

Due to the sample size and quality of the real-world medical data, an individual learner tends to either easily underfit or overfit. Ensemble learning, as a machine learning strategy that combines predictions from multiple base models, exhibits robust performance in prediction and classification tasks within the medical domain [10, 17]. In response to the characteristics of medical data, ensemble learning demonstrates superior predictive accuracy, enhanced model robustness, reduced risk of overfitting, and improved generalization capabilities compared to single models [10, 18, 19]. Therefore, we aimed to develop an ensemble learning model for predicting the risk of postoperative sepsis on a multicenter (Multi) dataset and conduct external validation on the Medical Information Mart for Intensive Care IV (MIMIC-IV) dataset with routine preoperative variables based on the better ones among eight conventional single models by comparing the model performance; and to investigate the important variables associated with postoperative sepsis.

Material and methods

This study involved model construction, internal validation and external validation. For model construction and internal validation, this multicenter study was conducted in three academic hospitals in different areas of China (Southwest Hospital of Third Military Medical University, Xuan Wu Hospital of capital medical university, and West China Hospital of Sichuan University). Ethical approvals were obtained for this retrospective study from the three hospitals (Certification Number: KY201936, 2019-132, 2021-349, respectively). For external validation, Medical Information Mart for Intensive Care IV (MIMIC-IV) [20] was incorporated, which is approved by the institutional review boards of Beth Israel Deaconess Medical Center in Boston, Massachusetts, and the Massachusetts Institute of Technology. Data were obtained from PhysioNet (https: //www.physionet.org/) by an author (Chunyong Yang, certification number: 46086293) with data usage agreement. No individual patient informed consent was required.

Patients

In the Multi dataset, clinical information of 49 768 surgical patients from the aforementioned hospitals between May 2014 and January 2020 were collected. The inclusion criteria were as follows: older than 18 years; American Society of

Anesthesiologists (ASA) score 2–4; undergoing abdominal surgery (spleen, gastrointestinal, hepatobiliary and pancreas, adrenal gland, urinary female reproductive); no sepsis, infection, or other serious complications before surgery. The exclusion criteria were as follows: undergoing surgery with local anesthesia; superficial or non-intra-abdominal surgical site; patients with multiple or incomplete surgical records; missing values > 30%. Herein, postoperative sepsis is determined as the presence of sepsis, severe sepsis, or septic shock after surgery by ICD-10 [21]. Patients with postoperative sepsis were determined as positive cases, while patients who met the criteria without postoperative sepsis or other serious postoperative complications were determined as negative cases. Due to the requirement of the algorithms, the negative cases were randomly matched according to the age range and surgical type of the positive cases with the ratio of 1 : 2.

For the validation dataset, MIMIC-IV includes information on 383 220 patients at Beth Israel Deaconess Medical Center from 2008 to 2019. The inclusion criteria were as follows: older than 18 years; ASA score 2–4; undergoing abdominal surgery and transformed into PACU; no infection or not consistent with the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) [1] before surgery; without important variables missing. The exclusion criteria were the same as for the Multi dataset. Postoperative sepsis was defined as fulfilling the Sepsis-3 criteria only after surgery within 30 days, and those who did not meet Sepsis-3 throughout hospitalization and had no diagnosis of sepsis were negative cases. Additionally, manual checking was performed according to the diagnosis after data extraction.

Data collection and processing

The whole process of the current study included data pre-processing, feature selection, model fitting and evaluation (Figure 1). Continuous variables were processed and standardized by normalization. After data pre-processing, the Multi dataset was randomly split into training (70%) and test (30%) datasets according to whether diagnosed with postoperative sepsis or not. Then feature selection and model fitting were conducted on the training dataset, while model evaluation was conducted on the test dataset for internal validation, and the MIMIC dataset for external validation.

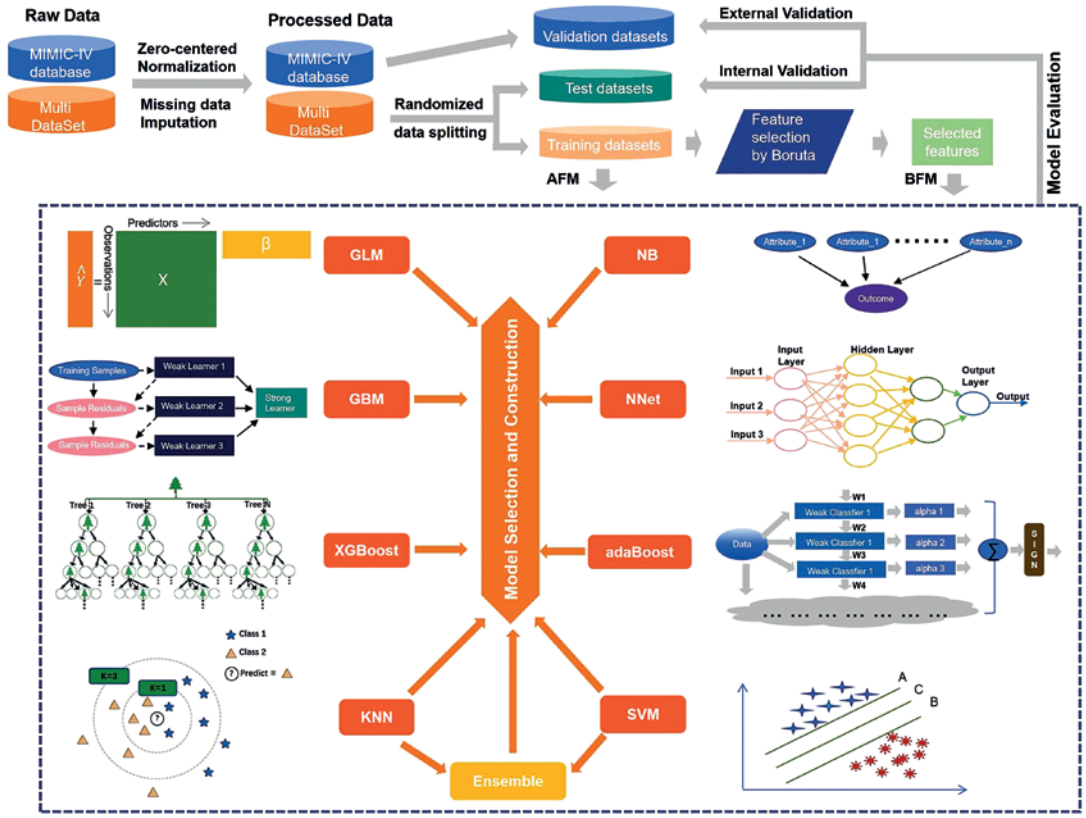

Figure 1. Development flow from raw data to model fitting and evaluation

GBM – Gradient Boosting Machine, GLM – Generalized Linear Models, KNN – k-Nearest Neighbor, XGBoost – Extreme Gradient Boosting, NNET – Neural Network, SVM – Support Vector Machine, AdaBoost – Adaptive Boosting, NB – Naive Bayes, AFM – models based on all variables, BFM – models based on the variables selected by Boruta.

The preoperative clinical variables were selected based on the accessibility and the association with the progress of postoperative sepsis according to the experience of the physicians in the three academic hospitals. The laboratory results closest to the surgical day were used. The abdominal procedures were classified by the procedure approach and surgical site. Finally, the following features were extracted: (1) demographic features, including age, body mass index (BMI), sex; (2) comorbidities: chemotherapy, hypertension, diabetes, cardiopathy, chronic obstructive pulmonary diseases (COPD), nephropathy, cancer; (3) laboratory parameters: albumin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), bilirubin, K^+, creatinine, glucose, white blood cells (WBC), platelets, hemoglobin, neutrophils, lymphocytes, neutrophil-lymphocyte ratio (NLR); (4) surgical information: emergency surgery, ASA score, type of surgery, procedure site.

Missing data that occur in more than one variable present a special challenge. The patients with missing values more than 30% were excluded, while others were imputed with KNN through DMwR2 R Packages [22]. The distribution of variables before and after imputation was assessed to ensure consistency.

Model construction and importance ranking

Two types of models were constructed based on feature-selected variables and all variables respectively. Feature selection on the training dataset was completed with the Boruta algorithm [23]; when the median variable importance in the set runs was significantly higher or lower than the median of the maximum values for the shadow attribute (blue), the variable was confirmed as important (green) or rejected as unimportant (red); otherwise it was tentatively important (yellow). For comparison, we implemented eight conventional ML algorithms: Gradient Boosting Machine (GBM), Generalized Linear Models (GLM), KNN, Extreme Gradient Boosting (XGBoost), Neural Network (NNET), SVM, Adaptive Boosting (AdaBoost), and Naive Bayes (NB). The ensemble learning model was constructed from two individual models with better performance. The models based on the variables selected by Boruta were labelled the BFM group, while those based on all variables were labelled the AFM group. Meanwhile, to investigate the important variables associated with postoperative sepsis, importance ranking was conducted on the training dataset by the final model.

Model evaluation

The model evaluation process was conducted using R (version 4.2.2). After the construction, the models' performance was evaluated by the area under curves (AUCs) of the receiver operating characteristic curves (ROC), and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy (ACC, the arithmetic means of sensitivity and specificity) and F1 score (the harmonic means of the precision and recall). Through the comparisons of the above metrics, the better two models were chosen to develop an ensemble learning model. Finally, the ensemble learning model was evaluated and compared with the single model through the net reclassification index (NRI), integrated discrimination improvement (IDI) and decision curve analysis (DCA), which were used to compare the models' clinical benefits and utility [24–27]. Meanwhile, the calibration curve was also used to evaluate the models' degree of calibration. Confidence intervals (CI) for model performance metrics were generated by bootstrap resampling of each dataset.

Statistical analysis

Continuous variables were expressed as mean with standard deviation (SD) or median with interquartile ranges (IQR) according to the data distribution. The categorical variables were expressed as frequency and percentages. The difference in clinical variables between positive and negative patients was compared by *t* test, ANOVA, Mann-Whitney test, or non-parametric test for continuous data, and the χ^2 or Fisher exact test for categorical data. Two-tailed tests were employed throughout. $P < 0.05$ was considered to indicate statistical significance. All the analyses were performed using SPSS (version 26.0, IBM) and R (version 4.2.2).

Results

Descriptive characteristics

As shown in Figure 2, a total of 955 patients were included in our study (648 patients from the Multi dataset, 307 patients from the MIMIC dataset), of which 285 patients were diagnosed with postoperative sepsis. The comparisons of the general characteristics and preoperative laboratory results between the sepsis and non-sepsis group in Multi and MIMIC datasets are shown in Tables I and II, respectively. In these two datasets, patients in the sepsis group had older age, more comorbidities (cardiopathy, COPD and nephropathy), higher ASA score, higher creatinine, glucose, WBC, NLR, and lower albumin, hemoglobin, lymphocyte count, and frequency of emergency procedures, compared with those in the non-sepsis group (all $p < 0.05$).

There are 9 variables with missing values after excluding the patients with missing data exceeding 30%. Among them, 8 were preoperative laboratory results (Supplementary Table SI). No significant

difference between imputed and original data was found, with the distribution roughly the same (Supplementary Table SII). Patient demographics and characteristics between the Multi dataset (training dataset and test dataset) and MIMIC dataset (validation dataset) are presented in Supplementary Table SIII. The distribution was roughly the same between training and test datasets.

Model performance

As shown in Table III, the eight conventional machine learning models achieved relatively good performance in both the BFM and AFM groups on the internal validation (test) dataset, slightly inferior on the external validation dataset. All the models presented a decreasing trend from the test dataset to the validation data set. In this situation, in the validation dataset, only KNN and SVM have the ACC more than 65% in both the BFM and AFM groups. For BFM, KNN achieved the AUC of 0.758 (95% CI: 0.700–0.817) and ACC of 69.4% on the validation dataset. SVM yielded 0.761 (95% CI: 0.698–0.824) and ACC of 67.1% on the validation dataset. In the AFM group, KNN achieved the AUC of 0.744 (95% CI: 0.685–0.803) and the ACC

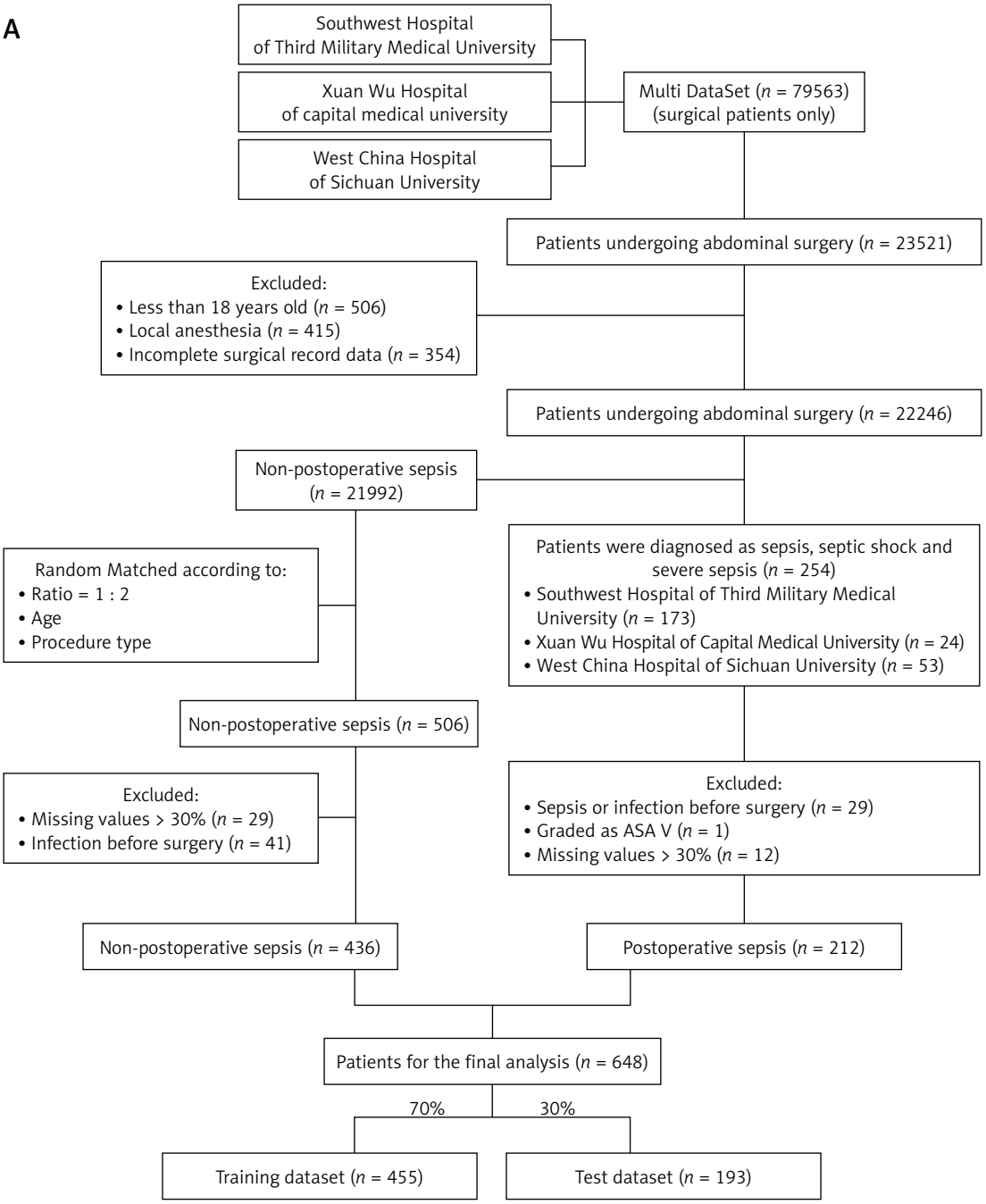

Figure 2. Consort diagram of the patient population. **A** – Multi dataset

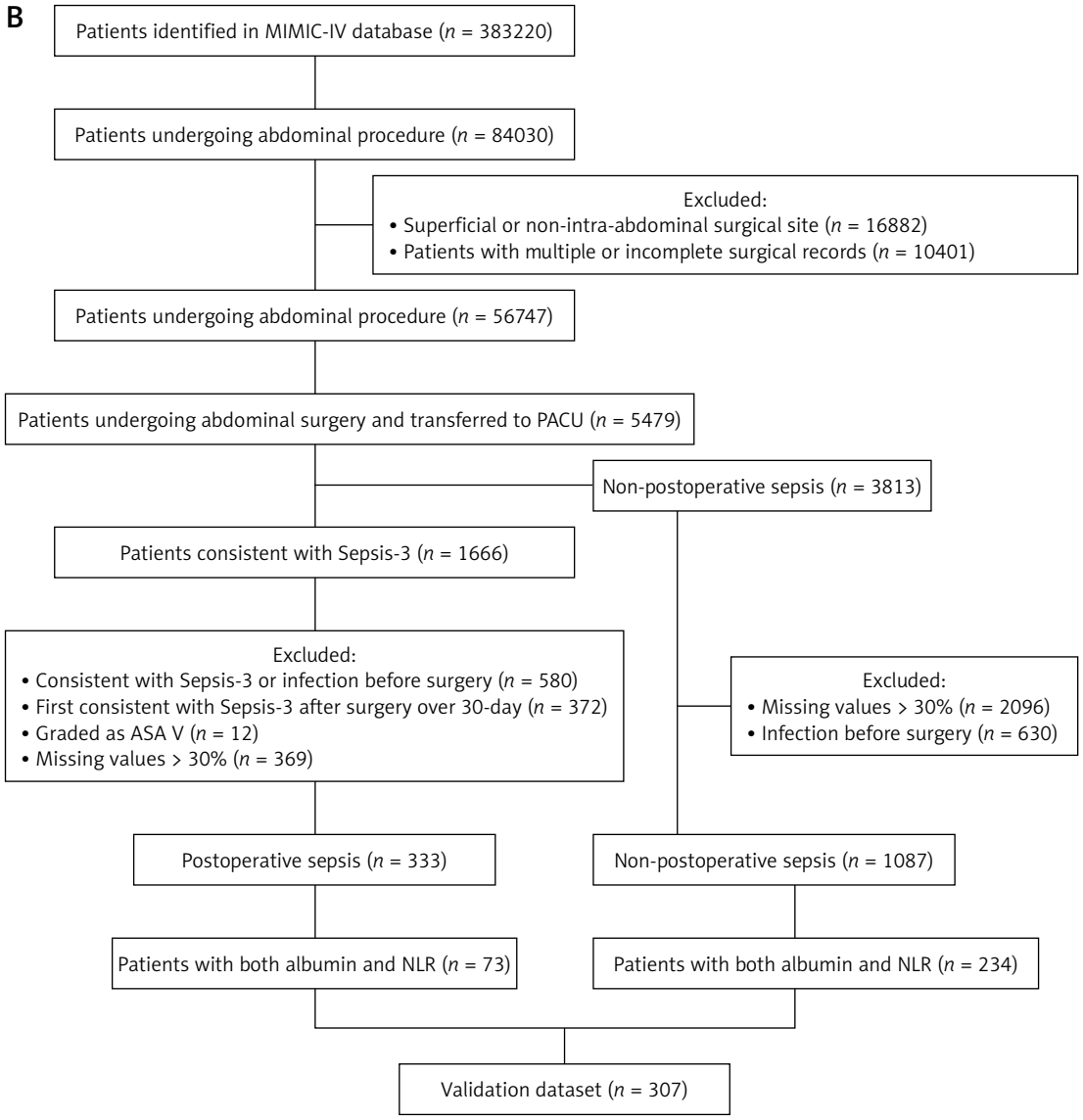

Figure 2. Cont. **B** – MIMIC-IV dataset

of 70.7% on the validation dataset. SVM yielded 0.768 (95% CI: 0.706–0.830) and the ACC of 66.4% on the validation dataset. Meanwhile, KNN and SVM yielded the specificity of more than 65% in BFM. In AFM, KNN also has specificity over 65% and SVM has the specificity of 64.1% (Figure 3).

The ensemble learning model was constructed by creating a linear blend of KNN and SVM. In BFM, the ensemble learning model yielded the AUCs of 0.892 (95% CI: 0.841–0.944) and 0.782 (95% CI: 0.727–0.838) on the test and validation dataset, respectively. In AFM, the AUCs were 0.877 (95% CI: 0.822–0.931) and 0.772 (95% CI: 0.713–0.831).

The NRI, IDI and DCA were used to compare the clinical benefits and utility among KNN, SVM and ensemble learning models. The ensemble learning model in BFM performed better than that in AFM; the NRI and IDI values were 0.064 and 0.023 on the test dataset, 0.846 and 0.585 on the validation dataset, respectively. In addition, in the BFM group, compared with the single model (KNN and SVM), the NRI values of the ensemble learning model were 0.080 and 0.215 on the test dataset, 0.002 and 0.036 on the validation dataset, while the IDI values were 0.016 and 0.090 on the test dataset, 0.020 and 0.033 on the validation dataset. As shown in Figure 4, the ensemble learning model in the BFM group performed better on calibration than the other models. In addition, when the threshold is 50%, the ensemble learning model performs better in the net benefit on both test and validation datasets. It indicated that the ensemble learning model had greater accuracy and clinical utility than the single models or those based on all features.

Table I. Baseline information of patients with or without postoperative sepsis in Multi dataset

Multi dataset	Sepsis (*n* = 212)	Non-sepsis (*n* = 436)	*P*-value
Age [years] M (P25, P75)	63.50 (48.0, 75.0)	52.0 (38.0, 66.0)	< 0.001
BMI [kg/m^2] M (P25, P75)	22.88 (20.70, 25.14)	23.05 (20.55, 25.69)	0.582
Female, *n* (%)	90 (42.5)	220 (50.5)	0.056
Comorbidity, *n* (%)			
Chemotherapy	14 (6.6)	1 (0.2)	< 0.001
Hypertension	51 (24.1)	55 (12.6)	< 0.001
Diabetes	31 (14.6)	43 (9.9)	0.074
Cardiopathy	60 (28.3)	21 (4.8)	< 0.001
COPD	9 (4.3)	7 (1.6)	0.042
Nephropathy	43 (20.3)	27 (6.2)	< 0.001
Cancer	91 (42.9)	221 (50.7)	0.063
Emergency surgery, *n* (%)			< 0.001
Yes	132 (62.3)	182 (41.7)	
No	80 (37.7)	254 (58.3)	
ASA, *n* (%)			< 0.001
1	1 (0.5)	15 (3.4)	
2	82 (38.7)	330 (75.7)	
3	107 (50.5)	90 (20.6)	
4	22 (10.4)	1 (0.2)	
Type of surgery, *n* (%)			< 0.001
Open	157 (74.1)	400 (91.7)	
Laparoscopy	55 (25.9)	36 (8.3)	
Procedure site, *n* (%)			0.108
Spleen	0 (21.2)	1 (12.8)	
Gastrointestinal	148 (42.5)	275 (42.9)	
Hepatobiliary and pancreas	19 (9.0)	60 (13.8)	
Adrenal gland	11 (16.0)	12 (20.0)	
Urinary	0 (6.1)	1 (7.8)	
Female reproductive	34 (5.2)	87 (2.8)	
Preoperative laboratory results			
Albumin [g/dl] M (P25, P75)	3.31 (2.81, 3.83)	4.11 (3.90, 4.37)	< 0.001
AST [IU/l] M (P25, P75)	18.40 (11.95, 33.0)	18.50 (12.90, 29.85)	0.773
ALT [IU/l] M (P25, P75)	25.35 (18.28, 44.18)	23.0 (18.0, 32.0)	0.012
Bilirubin [mg/dl] M (P25, P75)	0.85 (0.55, 1.46)	0.71 (0.52, 0.98)	< 0.001
K^+ [mmol/l] M (P25, P75)	3.94 (3.68, 4.26)	4.03 (3.76, 4.28)	0.079
Creatinine [mg/dl] M (P25, P75)	0.84 (0.67, 1.17)	0.74 (0.60, 0.89)	< 0.001
Glucose [mg/dl] M (P25, P75)	121.68 (97.02, 157.32)	97.92 (84.60, 120.06)	< 0.001
WBC [× 10^9/l] M (P25, P75)	7.79 (5.42, 11.71)	6.86 (5.39, 8.91)	0.002
Platelets [× 10^9/l] M (P25, P75)	185.0 (138.25, 242.75)	194.0 (148.0, 257.0)	0.070
Hemoglobin [g/dl] M (P25, P75)	11.85 (10.10, 13.38)	12.70 (11.40, 13.80)	< 0.001
Neutrophils [× 10^9/l] M (P25, P75)	5.64 (3.62, 9.41)	4.61 (3.35, 6.50)	< 0.001
Lymphocytes [× 10^9/l] M (P25, P75)	1.17 (0.73, 1.62)	1.35 (1.03, 1.74)	< 0.001
NLR [M (P25, P75)	5.33 (2.67, 11.22)	3.28 (2.23, 5.24)	< 0.001

Data are expressed as number (proportion), median (IQR [range]). BMI – body mass index, COPD – chronic obstructive pulmonary diseases, ALT – alanine aminotransferase, AST – aspartate aminotransferase, WBC – white blood cells, NLR – neutrophil-lymphocyte ratio.

Table II. Baseline information of patients with or without postoperative sepsis in MIMIC-IV dataset

MIMIC-IV dataset	Sepsis (*n* = 73)	Non-sepsis (*n* = 234)	*P*-value
Age [years] (mean, SD)	64.60 (1.29)	59.12 (1.19)	0.002
BMI [kg/m^2] M (P25, P75)	27.22 (23.64, 34.18)	28.30 (22.66, 30.98)	0.266
Female, *n* (%)	36 (49.3)	142 (60.7)	0.086
Comorbidity, *n* (%)			
Chemotherapy	8 (11.0)	14 (6.0)	0.150
Hypertension	0	6 (2.6)	0.369
Diabetes	29 (39.7)	46 (19.7)	< 0.001
Cardiopathy	42 (57.5)	51 (21.8)	< 0.001
COPD	15 (20.5)	13 (5.6)	0.042
Nephropathy	56 (76.7)	53 (22.6)	< 0.001
Cancer	13 (17.8)	35 (15.0)	0.558
Emergency surgery, *n* (%)			
Yes	44 (60.3)	102 (43.6)	0.013
No	29 (39.7)	132 (56.4)	
ASA, *n* (%)			
1	2 (2.7)	70 (33.3)	< 0.001
2	46 (63.0)	120 (51.3)	
3	20 (27.4)	30 (12.8)	
4	5 (6.8)	6 (2.6)	
Type of surgery, *n* (%)			< 0.001
Open	13 (17.8)	134 (57.3)	
Laparoscopy	60 (82.2)	100 (42.7)	
Procedure site, *n* (%)			0.021
Spleen	0	2 (0.9)	
Gastrointestinal	57 (78.1)	117 (50.0)	
Hepatobiliary and pancreas	13 (17.8)	102 (43.6)	
Adrenal gland	0	0	
Urinary	3 (4.1)	7 (3.0)	
Female reproductive	0	6 (2.6)	
Preoperative laboratory results			
Albumin [g/dl] (mean, SD)	3.08 (0.08)	3.69 (0.48)	< 0.001
AST [IU/l] M (P25, P75)	37 (20.5, 57.5)	27 (16, 43)	0.676
ALT [IU/l] M (P25, P75)	23 (13, 48.5)	23 (14, 42)	0.076
Bilirubin [mg/dl] M (P25, P75)	0.6 (0.4, 1.8)	0.6 (0.3, 0.9)	0.468
K^+ [mmol/l] M (P25, P75)	4 (3.7, 4.4)	4 (3.8, 4.2)	0.952
Creatinine [mg/dl] M (P25, P75)	1.1 (0.7, 2.05)	0.7 (0.6, 0.8)	< 0.001
Glucose [mg/dl] M (P25, P75)	125.0 (96.0, 159.0)	103.0 (90.0, 125.0)	< 0.001
WBC [× 10^9/l] M (P25, P75)	10.30 (6.40, 14.90)	8.2 (6.2, 12.0)	0.044
Platelets [× 10^9/l] (mean, SD)	193.47 (11.82)	241.73 (5.66)	< 0.001
Hemoglobin [g/dl] M (P25, P75)	7.95 (7.95, 9.6)	8.8 (8.8, 12.2)	< 0.001
Neutrophils [× 10^9/l] M (P25, P75)	8.54 (4.47, 13.13)	7.34 (4.31, 12.48)	0.189
Lymphocytes [× 10^9/l] M (P25, P75)	0.98 (0.55, 1.42)	1.04 (0.75, 2.75)	0.003
NLR, M (P25, P75)	7.54 (4.11, 15.64)	6.95 (3.71, 12.7)	0.005

Data are expressed as number (proportion), mean (SD) or median (IQR [range]). BMI – body mass index, COPD – chronic obstructive pulmonary diseases, ALT – alanine aminotransferase, AST – aspartate aminotransferase, WBC – white blood cells, NLR – neutrophil-lymphocyte ratio, SD – standard deviation.

Table III. Performance of models based on different algorithms in BFM and AFM group

Datasets	AUC (95% CI)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	F1 score	ACC (%)
Models based on selected variables by Boruta							
Model 1: GBM							
Test	0.886 (0.835–0.936)	68.3	91.5	79.6	85.6	0.735	83.9 (77.9–88.8)
Validation	0.762 (0.705–0.818)	90.4	49.2	35.7	94.3	0.512	59.0 (53.2–64.5)
Model 2: GLM							
Test	0.876 (0.821–0.932)	66.7	89.2	75.0	84.7	0.706	81.9 (75.7–87.0)
Validation	0.737 (0.668–0.805)	80.8	55.1	36.0	90.2	0.498	61.2 (55.5–66.7)
Model 3: KNN							
Test	0.860 (0.803–0.918)	55.6	96.2	87.5	81.7	0.680	82.9 (76.8–87.9)
Validation	0.758 (0.700–0.817)	74.0	68.0	41.9	89.3	0.535	69.4 (63.9–74.5)
Model 4: XGBoost							
Test	0.873 (0.815–0.931)	61.9	92.3	79.6	83.3	0.696	82.4 (76.3–87.5)
Validation	0.747 (0.688–0.806)	87.7	46.2	33.7	92.3	0.487	56.0 (50.3–61.7)
Model 5: NNET							
Test	0.876 (0.821–0.931)	74.6	83.1	68.1	87.1	0.712	80.3 (74.0–85.7)
Validation	0.747 (0.681–0.813)	84.9	45.7	32.8	90.7	0.473	55.1 (49.3–60.7)
Model 6: SVM							
Test*	0.877 (0.820–0.934)	60.3	92.3	79.2	82.8	0.685	81.9 (75.7–87.0)
Validation	0.761 (0.698–0.824)	74.0	65.0	39.7	88.9	0.517	67.1 (61.5–72.3)
Model 7: AdaBoost							
Test	0.877 (0.820–0.934)	61.9	90.0	75.0	83.0	0.678	80.8 (74.6–86.1)
Validation	0.761 (0.705–0.818)	86.3	51.7	35.8	92.4	0.506	59.9 (54.2–65.5)
Model 8: NB							
Test	0.863 (0.805–0.920)	74.6	83.1	68.1	87.1	0.712	80.3 (74.0–85.7)
Validation	0.685 (0.610–0.760)	82.2	22.7	24.9	80.3	0.382	36.8 (31.4–42.5)
Model 9: Ensemble Learning (KNN + SVM)							
Test*	0.892 (0.841–0.944)	65.1	94.6	85.4	84.8	0.739	84.9 (79.1–89.7)
Validation†	0.782 (0.726–0.838)	76.7	65.4	40.9	90.0	0.533	68.1 (62.5–73.3)
Models based on all variables							
Model 1: GBM							
Test	0.891 (0.842–0.939)	65.1	91.5	78.9	84.4	0.713	82.9 (76.8–87.9)
Validation	0.745 (0.686–0.805)	86.3	50.4	35.2	92.2	0.500	59.0 (53.2–64.5)
Model 2: GLM							
Test	0.876 (0.821–0.932)	66.7	89.2	75.0	84.7	0.706	81.9 (75.7–87.0)
Validation	0.737 (0.669–0.805)	79.5	54.7	35.4	89.5	0.489	60.6 (54.9–66.1)
Model 3: KNN							
Test	0.805 (0.738–0.872)	42.9	95.4	81.8	77.5	0.563	78.2 (71.7–83.8)
Validation	0.744 (0.685–0.803)	67.1	71.8	42.6	87.5	0.521	70.7 (65.3–75.7)
Model 4: XGBoost							
Test	0.870 (0.816–0.923)	63.5	90.0	75.5	83.6	0.690	81.4 (75.1–86.6)
Validation	0.774 (0.717–0.831)	83.6	53.0	35.7	91.2	0.500	60.3 (54.5–65.8)
Model 5: NNET							
Test	0.876 (0.822–0.929)	74.6	83.9	69.1	87.2	0.718	80.8 (74.6–86.1)
Validation	0.778 (0.718–0.838)	94.5	42.7	34.0	96.2	0.500	55.1 (49.3–60.7)

Table III. Cont.

Models based on selected variables by Boruta							
Datasets	**AUC (95% CI)**	**Sens (%)**	**Spec (%)**	**PPV (%)**	**NPV (%)**	**F1 score**	**ACC (%)**
Model 6: SVM							
Test	0.873 (0.817–0.928)	61.9	93.9	83.0	83.6	0.709	83.4 (77.4–88.4)
Validation	0.768 (0.706–0.830)	74.0	64.1	39.1	88.8	0.512	66.4 (60.9–71.7)
Model 7: AdaBoost							
Test	0.873 (0.817–0.928)	58.7	90.0	74.0	81.8	65.487	79.8 (73.4–85.2)
Validation	0.779 (0.725–0.834)	80.8	56.8	36.9	90.5	50.644	62.5 (56.9–68.0)
Model 8: NB							
Test	0.858 (0.801–0.915)	73.0	80.8	64.8	86.1	68.657	78.2 (71.7–83.8)
Validation	0.709 (0.639–0.780)	86.3	23.5	26.0	84.6	40.000	38.4 (32.7–44.1)
Model 9: Ensemble Learning (KNN + SVM)							
Test	0.877 (0.822–0.931)	58.7	94.6	84.1	82.6	69.159	82.9 (76.8–87.9)
Validation†	0.772 (0.713–0.831)	76.7	65.8	41.2	90.1	53.589	68.4 (62.9–73.6)

*GBM – Gradient Boosting Machine, GLM – Generalized Linear Models, KNN – k-Nearest Neighbor, XGBoost – Extreme Gradient Boosting, NNET – Neural Network, SVM – Support Vector Machine, AdaBoost – Adaptive Boosting, NB – Naive Bayes; *ensemble learning vs. SVM in models based on selected variables by Boruta on the test dataset, NRI and IDI were p < 0.05; †ensemble learning in models based on the variables selected by Boruta vs ensemble learning in models based on all variables on the validation dataset; NRI and IDI were p < 0.05.*

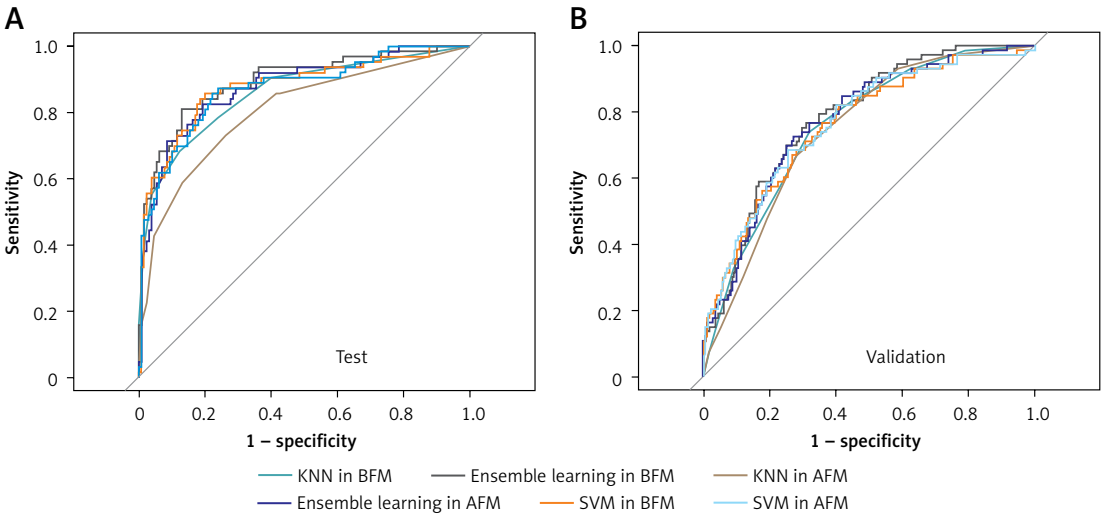

Figure 3. ROCs of SVM, KNN, and ensemble learning models in test and validation datasets

KNN – k-Nearest Neighbor, SVM – Support Vector Machine, AFM – models based on all variables, BFM – models based on the variables selected by Boruta.

Feature selection and importance analysis

As shown in Figure 5 A, albumin, cardiopathy, ASA score, type, age, neutrophil, creatinine, NLR, bilirubin, WBC, chemotherapy, glucose, lymphocyte count, emergency, hemoglobin, cancer, and AST were confirmed as important by Boruta. ALT, site, and BMI were found to be tentatively important. These variables were used for model construction and fitting, while others were rejected as unimportant.

Meanwhile, according to the results of each feature's contribution determined by KNN and SVM, the importance ranking of the selected variables in the actual model was also conducted (Figures 5 B, C). Albumin, ASA score, NLR, age and glucose were identified as the top-ranking features associated with postoperative sepsis.

Discussion

In this study, we developed models using ML algorithms based on routine variables from the Multi dataset to predict the risk of postoperative sepsis for patients undergoing abdominal surgery, and externally validated the models on the MIMIC IV dataset. It may help doctors preoperatively screen patients at high risk for postoperative sepsis, then provide timely management,

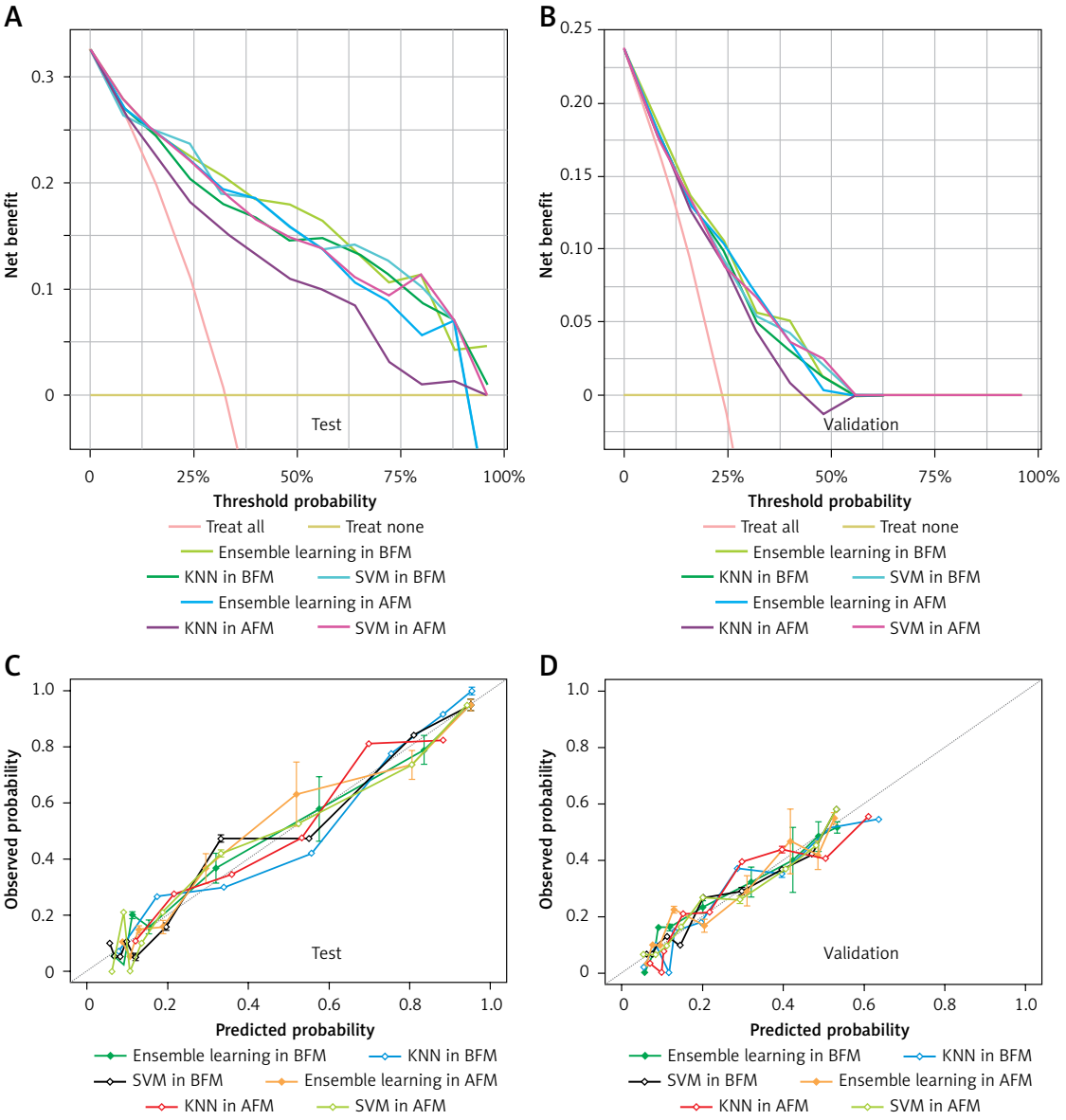

Figure 4. DCAs and calibration curves of SVM, KNN, and ensemble learning models in test and validation datasets

KNN – k-Nearest Neighbor, SVM – Support Vector Machine, AFM – models based on all variables, BFM – models based on the variables selected by Boruta.

eventually benefiting surgical patients. Lower albumin, higher ASA score, older age, higher NLR, and higher glucose were the important indicators associated with postoperative sepsis, suggesting that doctors should pay more attention to them preoperatively.

Early recognition is particularly important as prompt management of septic patients may improve outcomes [28]. The Implementation of National Early Warning Score (NEWS), one of the best early warning scores for sepsis in England, was confirmed to reduce mortality in the suspicion of sepsis cohort [29]. Moreover, Croft *et al.* reported that applying a computerized sepsis management system would increase early recognition by 12% and reduce hospital mortality by 6% [30].

With the growing attempts and endeavors for applying ML algorithms as new tools to solve medical problems, the quality of data, and the generalizability and robustness of models are the main limitations for application to real-world data. To some extent, the quality of data from the real world is the main problem; it includes the limited number of positive cases, heterogeneity of data, and numerous missing data. As we described above, the incidence of postoperative sepsis is low both in the current study and other existing ones, which may be one of the reasons that research on preoperatively screening patients at high risk of postoperative sepsis is limited. The heterogeneity of medical data is widely accepted, as the incidence of postoperative sepsis varies among medical cen-

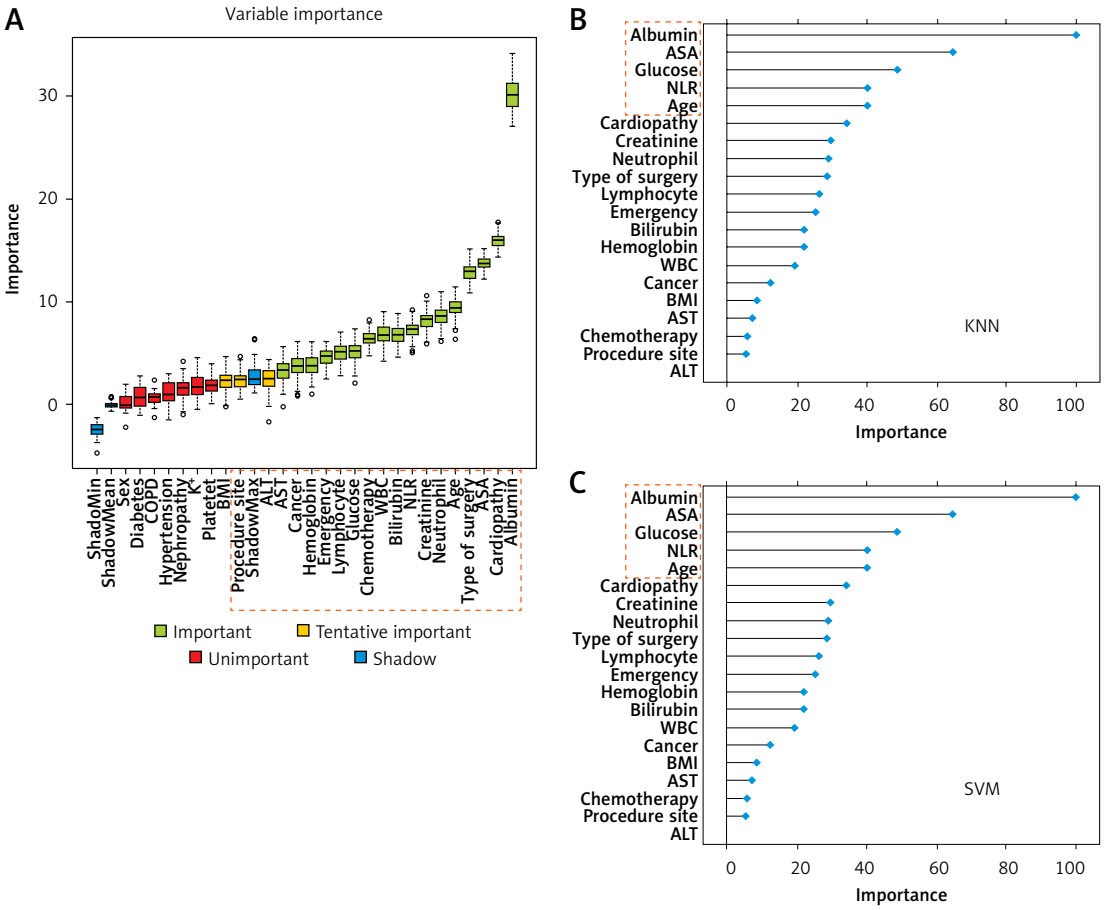

Figure 5. Results of feature selection and importance ranking. **A** – Feature selection by Boruta. The red box was rejected, the yellow one was tentative, and the green one was confirmed as important. Meanwhile, the rectangular box with red dotted line represents the variables selected for model construction. **B** – Feature importance ranking calculated by derivation from KNN on the training dataset. The rectangular box with red dotted line represents the top ranked features that were correlated with postoperative sepsis. **C** – Feature importance ranking calculated by derivation from SVM on the training dataset. The rectangular box with red dotted line represents the top ranked features that were correlated with postoperative sepsis

KNN – k-Nearest Neighbor, SVM – Support Vector Machine, BMI – body mass index, COPD – chronic obstructive pulmonary diseases, ALT – alanine aminotransferase, AST – aspartate aminotransferase, WBC – white blood cells, NLR – neutrophil-lymphocyte ratio.

ters and countries, and representative datasets are sorely needed to enhance the generalizability and robustness of models. However, most of the existing studies were conducted on single-center datasets or public datasets due to data availability and standardization [8–10]. To address the growing need, we made efforts to improve the robustness of our model via multi-source data. Our data were extracted from three representative large-scale academic hospitals in China. Moreover, most of the previous studies only performed internal validation [7, 8], which resulted in the unclear robust performance of the established models. In this study, we also used the MIMIC IV dataset as a validation dataset to perform external validation, which was quite different from our multi-source data. On the other hand, most real-world data have missing values, especially the multicenter studies, as well as MIMIC IV. Only a few studies using their own hospital data allow data sharing [31], which may aggravate the limitations on screening for postoperative sepsis. The lack of high-quality data and scarcity of models developed or validated in middle- or low-income countries are likely promoting inequality in healthcare. Herein, we excluded patients with more than 30% missing data to make the selected feature relatively integrated. Also, KNN was used to impute missing values for cases with less than 30% missing to increase volumes. By collecting the routine structured data and preprocessing the missing data, the data quality was increased, and our model may be applied in other centers. For increasing the robustness and generalizability of the model, the ML algorithms used were also important. Besides constructing single conventional ML models, we chose the best two to build an ensemble learning model through the caretEnsemble R package.

According to the results of NRI and IDI, the ensemble learning model was comparable to KNN and SVM. However, based on calibration curves and DCA curves, the ensemble learning model performed better than KNN and SVM. Ensemble learning could flexibly assemble prediction models to build an accurate one, which has been shown effective in many applications [32, 33]. Likewise, the relatively satisfactory model performance manifested in our study indicated its potential for real-world medical datasets.

Compared with other existing studies, we had several advantages. Studies for predicting postoperative sepsis are scarce. Among them, most were for a single surgical type, or based on single-center data, or not for preoperatively screening high-risk patients, or used non-routine variables, or had no external validation. Bunn *et al*. developed a tool to screen patients at high risk for postoperative sepsis based on LR, RF, XGboost and support vector machine (SVM) algorithms with a total of 223,214 appendectomies from the national surgery quality improvement program database (NSQIP). However, it achieved only moderate discrimination ability (a maximum AUC of 0.7) on a test dataset [7]. Zhang *et al.* developed a postoperative sepsis scoring tool for hepatobiliary and pancreatic surgery from a single-center dataset based on the LR algorithm, but, due to the unbalance of data (total patients: 522, postoperative sepsis: 55), the PPV was only 35% [8]. Moreover, some variables used for model development, such as interleukin and TNF-α, are not routine, which further affected the promotion of this tool [8, 15]. Our model was developed and evaluated on representative multicenter data, and variables were all preoperative and routine. Furthermore, the performance of our model was relatively satisfactory, with external validation in the MIMIC dataset acceptable. Therefore, our model has greater potential for application to other centers for preoperative screening of patients at high risk for postoperative sepsis.

Feature selection and importance ranking help elevate the model's performance and interpretability. In the current study, we chose the variables associated with the progression of postoperative sepsis based on the literature or experience of the physicians for the initial analysis. Then, feature selection was conducted using the Boruta algorithm, which is powerful, fast, and robust for both high-dimensional and low-dimensional datasets [23]. After model construction and fitting, we also ranked the importance of selected variables. Taking the rank by Boruta and importance ranking into consideration, the top 5 predictors were found to be the most important for postoperative sepsis. They were albumin, ASA score, NLR, age and glucose. Among these features, lower albumin, higher ASA score, and older age reflect the poor physiological state of the patient. Intact innate and adaptive immune responses depend on albumin, and low albumin is associated with increased risks of severity and death in patients with severe sepsis or organ failure [34]. NLR, a biomarker of systemic inflammation, indicates the balance between neutrophil and lymphocyte counts, and high NLR may indicate unfavorable prognoses in patients with sepsis [35, 36]. High glucose levels at sepsis onset have been proved to be independently associated with a worse prognosis, irrespective of the presence or absence of preexisting diabetes [37, 38]. Despite several features not actually being abnormal, it is recommended that anesthesiologists and surgeons pay more attention to them and adjust them to appropriate levels before surgery. The other variables selected for final analysis in our study, such as creatinine, neutrophil, WBC, ALT, and cardiopathy, are known to be clinically associated with postoperative sepsis [39–42]. Due to the important roles of albumin and NLR (these two variables were missing in more than 70% of the MIMIC-IV dataset), patients without albumin or NLR were excluded during data processing. Especially, anesthesiologists and surgeons could complete these two examinations before surgery, as both of them are routine and accessible.

There are a few limitations of this study. First, every effort was made to collect all patients with postoperative sepsis, but the amount of data is still relatively small compared with other big-data studies, and a larger volume of data is needed to improve the robustness. Second, this is a retrospective study, which may have the problems of missing data and inaccurate diagnosis and can only establish associations between factors, rather than causality. Prospective multicenter studies should be carried out to validate our model in the near future.

In conclusion, we confirm the feasibility of using an ensemble learning model based on KNN and SVM to accurately predict postoperative sepsis in patients undergoing abdominal surgery based on routine preoperative indicators. Meanwhile, albumin, ASA score, age, NLR, and glucose were considered as the important variables, suggesting that doctors should pay more attention to these variables preoperatively.

Funding

This study was supported by the National Key R&D Program of China (No. 2018YFC0116702 and No. 2018YFC0116704), National Science Foundation of China (No. 82070630 and No. 82100658), Special support for Chongqing postdoctoral research project in 2020 and Chongqing Talents Project (No. CQYC202103080).

Ethical approval

Approval number: KY201936, 2019-132, 2021-349.

Conflict of interest

The authors declare no conflict of interest.

References

1. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 2016; 315: 801-10.
2. Plaeke P, De Man JG, Coenen S, Jorens PG, De Winter BY, Hubens G. Clinical- and surgery-specific risk factors for post-operative sepsis: a systematic review and meta-analysis of over 30 million patients. Surg Today 2020; 50: 427-39.
3. Fried E, Weissman C, Sprung C. Postoperative sepsis. Curr Opin Crit Care 2011; 17: 396-401.
4. Bateman BT, Schmidt U, Berman MF, Bittner EA. Temporal trends in the epidemiology of severe postoperative sepsis after elective surgery: a large, nationwide sample. Anesthesiology 2010; 112: 917-25.
5. Brakenridge SC, Efron PA, Cox MC, et al. Current epidemiology of surgical sepsis: discordance between inpatient mortality and 1-year outcomes. Ann Surg 2019; 270: 502-10.
6. Bui MH, Khuong QL, Le PA, et al. Cost of postoperative sepsis in Vietnam. Sci Rep 2022; 12: 4876.
7. Bunn C, Kulshrestha S, Boyda J, et al. Application of machine learning to the prediction of postoperative sepsis after appendectomy. Surgery 2021; 169: 671-7.
8. Zhang H, Meng F, Lu S. Nomograms predicting the occurrence of sepsis in patients following major hepatobiliary and pancreatic surgery. Gastroenterol Res Pract 2020; 2020: 9761878.
9. Ren Y, Loftus TJ, Datta S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. JAMA Netw Open 2022; 5: e2211973.
10. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg 2019; 269: 652-62.
11. Wang XW, Niu XG, Li JX, Zhang SS, Jiao XF. SOFA score can effectively predict the incidence of sepsis and 30-day mortality in liver transplant patients: a retrospective study. Adv Ther 2019; 36: 645-51.
12. Peng Y, Zhang W, Xu Y, et al. Performance of SOFA, qSOFA and SIRS to predict septic shock after percutaneous nephrolithotomy. World J Urol 2021; 39: 501-10.
13. Niyongombwa I, Sibomana I, Karenzi ID, Kiswezi A, Rickard J. Kigali Surgical Sepsis (KiSS) score: a new tool to predict outcomes in surgical patients with sepsis in low- and middle-income settings. World J Surg 2020; 44: 3651-7.
14. Gabriel RA, Trivedi S, Schmidt UH. A point-based risk calculator predicting mortality in patients that developed postoperative sepsis. J Intensive Care Med 2021; 36: 1443-9.
15. Madushani R, Patel V, Loftus T, et al. Early biomarker signatures in surgical sepsis. J Surg Res 2022; 277: 372-83.
16. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020; 46: 383-400.
17. Zhong J, Zeng X, Cao W, et al. Semisupervised multiple choice learning for ensemble classification. IEEE Trans Cybern 2022; 52: 3658-68.
18. Luo S, Xu J, Jiang Z, et al. Artificial intelligence-based collaborative filtering method with ensemble learning for personalized lung cancer medicine without genetic sequencing. Pharmacol Res 2020; 160: 105037.
19. Horng H, Steinkamp J, Kahn CE Jr, Cook TS. Ensemble approaches to recognize protected health information in radiology reports. J Digit Imaging 2022; 35: 1694-8.
20. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.2). PhysioNet2023.
21. Bouza C, Lopez-Cuadrado T, Amate-Blanco JM. Use of explicit ICD9-CM codes to identify adult severe sepsis: impacts on epidemiological estimates. Crit Care 2016; 20: 313.
22. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min 2017; 10: 363-77.
23. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform 2019; 20: 492-503.
24. Uno H, Tian L, Cai T, Kohane IS, Wei LJ. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. Stat Med 2013; 32: 2430-42.
25. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006; 26: 565-74.
26. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA 2015; 313: 409-10.
27. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011; 30: 11-21.
28. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016; 315: 801.
29. Pullyblank A, Tavaré A, Little H, et al. Implementation of the National Early Warning Score in patients with suspicion of sepsis: evaluation of a system-wide quality improvement project. Br J Gen Pract 2020; 70: e381-e8.
30. Croft CA, Moore FA, Efron PA, et al. Computer versus paper system for recognition and management of sepsis in surgical intensive care. J Trauma Acute Care Surg 2014; 76: 311-7; discussion 8-9.
31. Brown SM, Jones J, Kuttler KG, Keddington RK, Allen TL, Haug P. Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department. BMC Emerg Med 2016; 16: 31.
32. Couvy-Duchesne B, Faouzi J, Martin B, et al. Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: ARAMIS contribution to the predictive analytics competition 2019 challenge. Front Psychiatry 2020; 11: 593336.
33. Zhu Y, Brettin T, Evrard YA, et al. Ensemble transfer learning for the prediction of anti-cancer drug response. Sci Rep 2020; 10: 18040.
34. Wiedermann CJ. Hypoalbuminemia as surrogate and culprit of infections. Int J Mol Sci 2021; 22: 4496.
35. Huang Z, Fu Z, Huang W, Huang K. Prognostic value of neutrophil-to-lymphocyte ratio in sepsis: a meta-analysis. Am J Emerg Med 2020; 38: 641-7.

36. Rehman FU, Khan A, Aziz A, Iqbal M, Mahmood SBZ, Ali N. Neutrophils to lymphocyte ratio: earliest and efficacious markers of sepsis. Cureus 2020; 12: e10851.
37. van Vught LA, Wiewel MA, Klein Klouwenberg PM, et al. Admission hyperglycemia in critically ill sepsis patients: association with outcome and host response. Crit Care Med 2016; 44: 1338-46.
38. Zohar Y, Zilberman Itskovich S, Koren S, Zaidenstein R, Marchaim D, Koren R. The association of diabetes and hyperglycemia with sepsis outcomes: a population-based cohort analysis. Intern Emerg Med 2021; 16: 719-28.
39. Shen XF, Cao K, Jiang JP, Guan WX, Du JF. Neutrophil dysregulation during sepsis: an overview and update. J Cell Mol Med 2017; 21: 1687-97.
40. Crouser ED, Parrillo JE, Seymour CW, et al. Monocyte distribution width: a novel indicator of sepsis-2 and sepsis-3 in high-risk emergency department patients. Crit Care Med 2019; 47: 1018-25.
41. Guarracino F, Bertini P, Pinsky MR. Cardiovascular determinants of resuscitation from sepsis and septic shock. Crit Care 2019; 23: 118.
42. Piotrowski D, Sączewska-Piotrowska A, Jaroszewicz J, Boroń-Kaczmarska A. Lymphocyte-to-monocyte ratio as the best simple predictor of bacterial infection in patients with liver cirrhosis. Int J Environ Res Public Health 2020; 17: 1727.