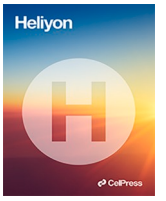
Heliyon 10 (2024) e34498

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Development and validation of a machine learning-based interpretable model for predicting sepsis by complete blood cell parameters

Tiancong Zhang [a,b,c], Shuang Wang [a,b,c], Qiang Meng [a,b,c], Liman Li [a,b,c], Mengxue Yuan [a,b,c], Shuo Guo [a,b,c], Yang Fu [a,b,c,*]

[a] Department of Laboratory Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China
[b] Sichuan Clinical Research Center for Laboratory Medicine, Chengdu, Sichuan, 610041, China
[c] Clinical Laboratory Medicine Research Center of West China Hospital, Chengdu, Sichuan, 610041, China

ARTICLE INFO

Keywords:
Sepsis
Machine learning
Complete blood cell parameters
Prediction model
Cell population data

ABSTRACT

Background: Sepsis, a severe infectious disease, carries a high mortality rate. Early detection and prompt treatment are crucial for reducing mortality and improving prognosis. The aim of this research is to develop a clinical prediction model using machine learning algorithms, leveraging complete blood cell (CBC) parameters, to detect sepsis at an early stage.

Methods: The study involved 572 patients admitted to West China Hospital of Sichuan University between July 2020 and September 2021. Among them, 215 were diagnosed with sepsis, while 357 had local infections. Demographic information was collected, and 57 CBC parameters were analyzed to identify potential predictors using techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest (RF), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost). The prediction model was built using Logistic Regression and evaluated for diagnostic specificity, discrimination, and clinical applicability including metrics such as the area under the curve (AUC), calibration curve, clinical impact curve, and clinical decision curve. Additionally, the model's diagnostic performance was assessed on a separate validation cohort. Shapley's additive explanations (SHAP), and breakdown (BD) profiles were used to explain the contribution of each variable in predicting the outcome.

Results: Among all the machine learning methods' prediction models, the LASSO-based model (λ = min) demonstrated the highest diagnostic performance in both the discovery cohort (AUC = 0.9446, $P < 0.001$) and the validation cohort (AUC = 0.9001, $P < 0.001$). Furthermore, upon local analysis and interpretation of the model, we demonstrated that LY-Z, MO-Z, and PLT-I had the most significant impact on the outcome.

Conclusions: The predictive model based on CBC parameters can be utilized as an effective approach for the early detection of sepsis.

* Corresponding author. Department of Laboratory Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China.
E-mail address: fuyang827@wchscu.cn (Y. Fu).

https://doi.org/10.1016/j.heliyon.2024.e34498
Received 17 April 2024; Received in revised form 25 May 2024; Accepted 10 July 2024
Available online 11 July 2024

Strengths of this study

1. This study is the first to construct a prediction model for early sepsis diagnosis using multiple blood cell parameters, including blood cell morphology-related parameters such as CPD and RDW.
2. The optimal diagnostic model was constructed using multiple machine learning algorithms, and model interpretation was performed.
3. In this study, we deeply excavated the CBC parameters from sepsis and common infection patients within 24 hours of admission, effectively identified several important characteristics of sepsis patients and constructed a clinical prediction model. The model may be greatly significant to judge the patient's condition and specify the treatment strategy in time.

1. Introduction

Sepsis remains a significant global public health threat, affecting over 30 million individuals annually. Tragically, 5 million people succumb to the disease or suffer lasting complications [1]. Prompt and accurate diagnosis is crucial for effective treatment and improved outcomes. However, the complexity and diversity of sepsis-related complications presents obstacles in the diagnostic process [2].

According to the Third International Consensus Definitions for Sepsis and Septic Shock, sepsis is defined as life-threatening organ dysfunction resulting from a dysregulated host response to infection. A Sequential [Sepsis-related] Organ Failure Assessment (SOFA) score of $\geq$2 is recommended for diagnosis [3]. However, due to the complexity of SOFA scoring, researchers have sought simpler, more direct markers associated with sepsis [4]. Thus far, in addition to the significance of procalcitonin [5], different neutrophil subsets have shown the ability to discriminate sepsis [6]. Certain cytokines, such as interleukin-1 receptor2 (IL-1R2), have been used to differentiate Gram-positive and Gram-negative bacterial infections [7].

Complete blood cell count (CBC) serves as the most common indicator for early suspicion of sepsis in clinical practice. Peripheral white blood cell (WBC) count and absolute neutrophil count have repeatedly proven valuable in the diagnosis and prognosis of sepsis [8-10]. In recent years, with advancements in CBC analyzers, an increasing number of blood cell parameters have presented good predictive value for sepsis. For instance, the monocyte distribution width (MDW) reported by the DxH 900 (Beckman Coulter Inc., USA) has been established as a screening tool for early identification of septic patients in large observational studies [11,12]. Additionally, cell population data (CPD) parameters reported by the XN analyzer (Sysmex Inc., Japan), which reflect the size and internal structure of leukocytes, also serve as good discriminators of sepsis [13,14].

Although numerous studies have shown the predictive value of CBC parameters in sepsis, the parameters related to blood cell morphogenesis based on cell size and particle complexity have not been fully explored and utilized. In this study, we aim to develop a sepsis predictive model based on CBC parameters, providing a laboratory foundation for early and accurate sepsis diagnosis.

2. Methods

2.1. Study population and variables collection

In the derivation cohort, we included 215 septic patients and 357 patients diagnosed with local infections who were admitted to West China Hospital of Sichuan University between July 2020 and September 2021. Demographic information, such as age, gender, primary infection sites, and blood culture positivity rate, was collected from the Hospital Information System. Furthermore, within 24 hours of admission, CBC parameters were gathered using the Sysmex-XN series (Sysmex, Japan) through the Laboratory Information Management System.

The inclusion criteria for the sepsis group were based on "The Third International Consensus Definitions for Sepsis and Septic Shock" (sepsis 3.0). The local infection group included patients with conditions such as pneumonia, acute appendicitis, cholangitis, urinary tract infection, skin infection, wound infection, and suppurative otitis media. Exclusion criteria comprised patients who died within 24 hours of enrollment, those who had used immunosuppressants or chemotherapy drugs within the last 3 months, and individuals with autoimmune diseases, malignant tumors, viral hepatitis, HIV infection, or tuberculosis.

Subsequently, we verified the diagnostic efficacy of the model in a separate cohort of 561 patients admitted to the emergency department at West China Hospital of Sichuan University between November 2021 and April 2024. Each patient displayed fever and exhibited symptoms of infection as observed clinically, through imaging tests, or laboratory findings. CBC data were collected within 24 hours of admission. After applying the aforementioned inclusion and exclusion criteria, the validation group comprised 207 individuals with sepsis and 354 patients with localized infections.

All procedures conducted in this study adhered to the principles of the Declaration of Helsinki and received approval from the Ethics Committee of West China Hospital (No. 2020-641).

2.2. Machine learning algorithm

Lasso regression (LASSO) is a technique employed to address overfitting and reduce the complexity of a linear regression model [15]. It is also valuable for selecting important characteristic variables. Regularization, a common technique in machine learning and statistical modeling, serves to mitigate the issue of overfitting in linear regression.

Random forest (RF) is an algorithm rooted in decision trees [16]. It trains decision trees on different samples and feature subsets,

Table 1
The clinical characteristic and baseline of the complete blood cell parameters.

Group	Derivation cohort(n = 572)		Validation cohort (n = 561)	
	sepsis(n = 215)	local infection(n = 357)	sepsis(n = 207)	local infection(n = 354)
age	56(43–70)	51(36–68)*	61(51–72)	53(35–68)*
sex(M/F)	138/77	190/167*	127/80	198/156*
primary infection sites, n(%)				/
respiratory infection	97(45.1 %)	231(64.7 %)	/	/
abdominal infection	81(37.7 %)	79(22.1 %)	/	/
skin infection	6(2.7 %)	9(2.5 %)	/	/
urinary tract infection	24(11.2 %)	18(5.1 %)	/	/
others	7(3.3 %)	20(5.6 %)	/	/
blood culture positive rate, n(%)	121(56.3 %)	/	/	/
complete blood cell parameters				
WBC(10^9/L)	8.88(6.04–12.65)	9.13(6.85–12.96)	9.5(6.65–13.27)	7.95(5.768–11.158)*
RBC(10^12/L)	3.26(2.73–3.99)	3.96(3.07–4.77)*	3.07(2.57–3.635)	4.16(3.605–4.63)*
HGB(g/L)	99(85–121)	112(91–133)*	89(75.5–104)	124(106–138)*
HCT(%)	31.2(27.0–37.8)	35.4(28.5–40.0)*	28.1(24.2–32.75)	38.4(33–42.77)*
MCV(fL)	96(92–102)	92(87–96)*	93.1(89.7–97.3)	91.9(88.53–95.38)
MCH(pg)	30.7(29.0–32.1)	30.2(28.3–31.4)*	29.4(28.5–30.7)	30.2(28.9–31.3)*
MCHC(g/L)	317(305–330)	323(313–333)*	316(306–326)	326.5(318–335)*
PLT(10^9/L)	137(80–223)	191(135–263)*	168(103–258)	229(175–297.8)*
RDW-SD(fL)	52.1(47.0–60.3)	46.4(42.3–53.3)*	49.7(45.7–55.45)	44.4(41.7–47.8)*
RDW-CV(%)	15.0(13.7–17.2)	14.2(13.0–16.3)*	14.6(13.6–16.9)	13.2(12.5–14.57)*
PDW(fL)	14.3(13.3–17.6)	13.5(11.3–15.0)*	14.8(12–17.2)	12.15(10.5–13.8)*
MPV(fL)	11.7(11.2–12.9)	11.2(10.1–11.9)*	11.8(10.7–12.9)	10.6(9.8–11.5)*
P-LCR(%)	37.9(33.8–47.6)	33.8(26.4–39.9)*	38.9(30–47.45)	29.6(23.12–36.4)*
PCT(%)	0.21(0.14–0.25)	0.23(0.17–0.27)*	0.23(0.15–0.29)	0.24(0.1925–0.31)*
NRBC#(10^9/L)	0.01(0–0.02)	0(0–0.01)*	0(0–0.01)	0(0–0)
NRBC%(%)	0.1(0.0–0.2)	0(0–0.1)*	0(0–0.1)	0(0-0)*
NEUT#(10^9/L)	7.05(4.69–10.66)	7.31(4.92–10.66)	7.54(5.285–11.5)	5.9(3.652–8.617)*
LYMPH#(10^9/L)	0.76(0.50–1.21)	1.03(0.65–1.46)*	0.78(0.49–1.255)	1.335(0.9325–1.82)*
MONO#(10^9/L)	0.48(0.28–0.70)	0.61(0.42–0.84)*	0.51(0.31–0.9)	0.56(0.43–0.83)
EO#(10^9/L)	0.02(0.00–0.09)	0.04(0.01–0.11)*	0.04(0–0.135)	0.08(0.02–0.17)
BASO#(10^9/L)	0.02(0.01–0.04)	0.02(0.01–0.04)	0.03(0.01–0.05)	0.03(0.02–0.04)
NEUT%(%)	83.4(74.6–88.8)	79.7(70.8–86.8)*	84.7(74.9–90)	74(62.85–80.97)*
LYMPH%(%)	9.0(5.5–15.1)	11.4(6.7–18.3)*	8.5(5.1–13.85)	16.7(10.7–24.5)*
MONO%(%)	5.7(3.5–8.0)	6.7(5.0–8.7)*	5.9(3.2–9.15)	7.3(6–8.975)*
EO%(%)	0.3(0.0–0.9)	0.5(0.1–1.5)*	0.5(0–1.5)	1.2(0.3–2.275)*
BASO%(%)	0.3(0.2–0.4)	0.3(0.2–0.4)	0.2(0.2–0.4)	0.35(0.2–0.5)*
IG#(10^9/L)	0.08(0.04–0.22)	0.06(0.03–0.15)*	0.1(0.05–0.195)	0.04(0.02–0.07)*
IG%(%)	1.0(0.5–1.9)	0.7(0.3–1.4)*	1.1(0.7–1.85)	0.5(0.3–0.8)*
PLT-I(10^9/L)	140(82–221)	195(132–270)*	167(101–260.5)	229(175–297.8)*
MicroR(%)	1.2(0.6–2.5)	1.3(0.8–3.4)*	1.8(1–2.55)	1.1(0.6–1.9)*
MacroR(%)	5.3(3.9–9.4)	3.9(3.4–5.3)*	3.8(3.1–6.05)	3.8(3.4–4.375)*
WBC-N(10^9/L)	8.88(6.09–12.65)	9.04(6.75–12.75)	9.5(6.65–13.27)	7.95(5.843–11.158)*
BA-N#(10^9/L)	0.02(0.01–0.04)	0.02(0.02–0.04)	0.03(0.01–0.05)	0.03(0.02–0.04)
BA-N%(%)	0.3(0.2–0.4)	0.3(0.2–0.4)	0.3(0.2–0.4)	0.4(0.2–0.6)*
WBC-D(10^9/L)	9.26(6.41–12.78)	9.57(7.05–12.88)	9.45(6.615–13.32)	8.035(5.86–11.08)*
HFLC#(10^9/L)	0.01(0–0.02)	0.01(0–0.02)	0.02(0.01–0.04)	0.01(0–0.02)*
HFLC%(%)	0.1(0–0.3)	0.1(0–0.3)	0.2(0.1–0.5)	0.1(0–0.2)*
BA-D#(10^9/L)	0.04(0.02–0.07)	0.03(0.02–0.04)*	0.03(0.01–0.04)	0.03(0.02–0.04)
BA-D%(%)	0.4(0.2–0.8)	0.3(0.2–0.5)*	0.3(0.1–0.5)	0.3(0.2–0.6)
NE-SSC	147.1(144.4–151.5)	153.5(150.4–156.8)*	151.5(146.8–155.8)	154.1(150.8–157.1)*
NE-SFL	54.0(48.6–58.8)	51.4(48.4–54.6)*	50.9(47.65–54.8)	48.45(46.1–51.58)*
NE-FSC	86.3(81.8–90.8)	90.0(86.9–93.5)*	86.3(82.7–90.25)	88.9(85.8–91.45)*
LY-X	83.1(81.0–84.5)	81.6(79.8–83.2)*	82.7(80.6–84.3)	81.35(78.92–83.1)*
LY-Y	67.9(63.6–72.3)	70.0(66.7–73.2)*	69.4(65.15–75.1)	68.6(65.6–72)
LY-Z	60.6(58.0–63.1)	57.4(55.7–59.0)*	59.1(57.55–60.85)	58.5(57.4–59.4)*
MO-X	121.7(118.9–124.3)	121.5(119.7–123.1)	122.6(120.8–124.7)	120.5(118.5–122.4)
MO-Y	112.8(102.0–121.5)	116.7(111.7–122.9)*	111.7(105.5–118.8)	112.8(107.3–118.5)
MO-Z	69.0(64.8–73.6)	65.8(62.5–69.4)*	66(63.45–68.9)	66.4(64.6–68.1)
NE-WX	356(329–381)	307(294–320)*	328(309–349.5)	303(292–314)*
NE-WY	731(661–796)	641(612–688)*	648(601–718.5)	612(586–638.8)*
NE-WZ	775(664–874)	618(584–660)*	695(607.5–763)	631.5(577.2–694)*
LY-WX	523(456–604)	492(445–527)*	513(472–577.5)	486.5(447.2–529.8)*
LY-WY	992(878–1140)	922(844–1013)*	942(854–1051.5)	889.5(830–951)*
LY-WZ	707(588–871)	525(489–573)*	581(489–712)	528(439.2–588)*
MO-WX	280(252–317)	257(241–276)*	264(243–290)	252(236–272)*
MO-WY	727(603–840)	693(628–765)	700(611.5–793)	691.5(622–758.8)
MO-WZ	644(573–747)	537(496–583)*	617(518–688.5)	573.5(505–634)*

Data are described as median with upper and lower quartiles for continuous variables. *compared with the sepsis group, $P < 0.05$.

then aggregates the prediction results of each tree through voting or average weighting to derive the final prediction results.

Support vector machine (SVM) is a boundary-based classification method that filters the most relevant features for the target variables [17]. This algorithm conducts variable screening through recursive feature elimination, iteratively removing features with the poorest model performance and retraining the model on the remaining features to obtain an optimal feature subset. This enables the model to achieve optimal performance on the selected subset.

eXtreme Gradient Boosting (XGBoost) is a gradient boosting algorithm used to solve classification and regression problems. It is based on decision tree models and utilizes a series of optimization strategies such as shrinkage, column subsampling, and row sampling to improve the generalization performance of the model.

2.3. Model explanation

Machine learning algorithms have demonstrated commendable performance; however, their lack of interpretability has constrained their applications, mainly attributable to their "black-box" nature. Consequently, enhancing the transparency and interpretability of models has become a prerequisite for the widespread adoption of machine-learning solutions. The SHAP (Shapley Additive exPlanations) model serves as an algorithm designed to interpret the predicted values of a machine learning model. The Breakdown (BD) explanation, on the other hand, constitutes a model-independent explanation method and falls under the category of feature attribution methods in local explanation.

2.4. Statistical analysis

Data with skewed distributions were presented using medians with quartiles (P25, P75), and Wilcoxon's tests were employed to compare groups. The R software (version 4.3.1) was utilized to implement machine learning algorithms. Candidate variables were screened using a machine learning algorithm to construct the regression model. The R packages used included "e1071", "glmnet", "pROC", "randomForest", "xgboost", "shapviz" and "ibreakdown". The model's overall performance was assessed using LR chi2, Pseudo R2, and *P* values. Prediction probability accuracy was measured using the Brier score, model differentiation was evaluated using the Kendall rank correlation coefficient (tau-a), and the model's specificity, sensitivity, negative predictive value (NPV), and positive predictive value (PPV) were assessed using the confusion matrix. Following a comprehensive evaluation of these indices, the

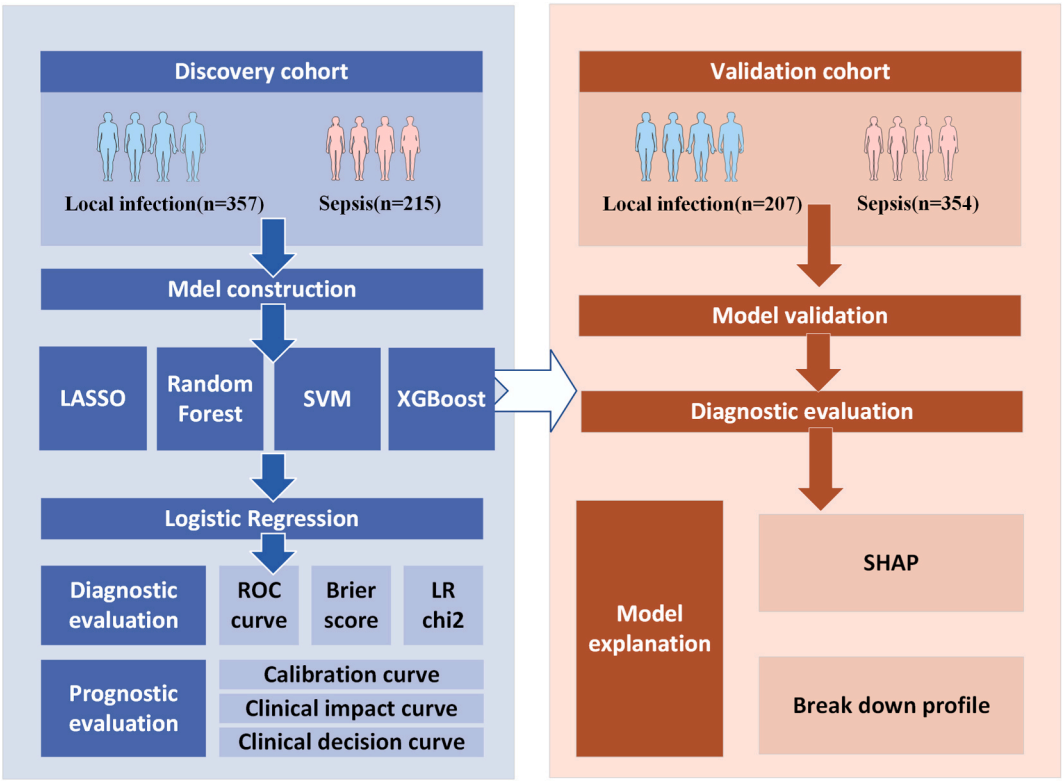


Fig. 1. An interpretable predictive diagnosis model of sepsis based on CBC parameters was constructed and validated by machine learning method.

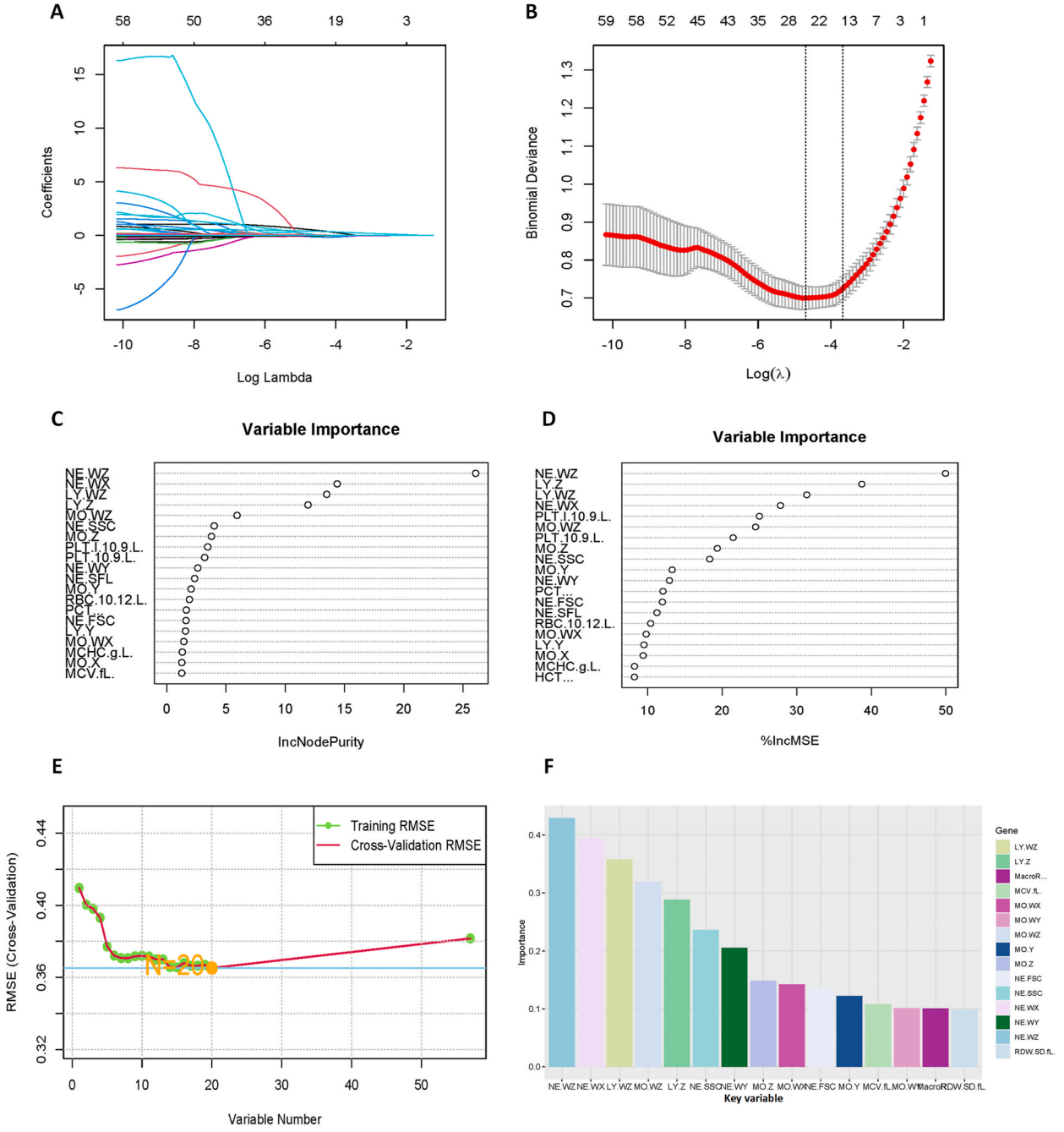


Fig. 2. Machine learning algorithm screening candidate variables. (2A): Trajectory diagram of variable coefficients. (2B): Cross-validation curve. The upper Horizontal axis represents the number of variables with non-zero coefficients, the lower Horizontal axis represents the logarithmic value of the penalty coefficient λ. The vertical axis represents the mean square error, with two vertical lines indicating λ_{min} and λ_{1se}. (2C): Variables are ranked by node purity. The Gini coefficient measures classification quality, while higher IncNodePurity values indicate greater importance. (2D): Variables are ranked by root mean square error at the split node. The impact of a feature is measured by (% IncMSE), which evaluates the increase in MSE after the feature is removed from the complete model. Higher (% IncMSE) values indicate more significant features. (2E): This graph shows how the number of features used for prediction affects the generalization error rate. The lowest error rate and highest accuracy are achieved when 20 characteristic variables are used. (2F): The top 15 variables of importance are ranked here.

model demonstrating the best performance was selected as the final diagnostic model.

3. Results

3.1. Patient characteristics

A total of 572 patients were included in the derivation cohort, among whom 215 patients were diagnosed with sepsis and 357 patients had local infections. The validation cohort consisted of 561 patients, comprising 207 septic patients and 354 patients with local infections. Demographic characteristics of all patients were collected, and 57 indicators related to CBC parameters, such as hematocyte count, leukocyte classification count, and morphological indicators associated with erythrocytes, platelets, and leukocytes, were gathered. A detailed comparison between the two groups is presented in Table 1.

3.2. Model construction and validation

The process of research is illustrated in Fig. 1. To select the most suitable variables, LASSO, RF, SVM, and XGBoost methods were employed.These machine learning methods are well-suited for handling binary classification and regression problems. Following this, a regression model was developed. Differentiation, specificity, and sensitivity of the model were assessed using various indicators. Finally, an external verification queue was introduced to evaluate the diagnostic effectiveness of the model.

The LASSO algorithm analysis resulted in 59 variables (57 CBC parameter,age and gender) being compressed as lambda (λ) increased, with unimportant variables compressed to zero (Fig. 2A). Two distinct lambda values were obtained, λ-min and λ-se, for the better and more concise models, respectively (Fig. 2B). The λ-min model gave better diagnostic performance with key variables being gender, age, PLT-I, NE-FSC, LY-Y, LY-Z, MO-Y, MO-Z, NE-WZ, and MO-WZ. The regression model's LR chi2 value, Pseudo R2, and *P* value were 442.87, 0.734, and $P < 0.0001$ respectively. Furthermore, the model offered high AUC at 0.9446 (95 % CI 0.9238–0.9654), specificity, sensitivity, NPV and PPV were 94.1 %, 83.3 %, 88.77 %, and 91.01 %, respectively (Table 2).

The random forest algorithm showed that root mean square error and Gini impurity had similar outcomes (Fig. 2C,D). The top 10 variables were identified, of which nine were common. The random forest regression model revealed that the unpurified Gini split node was more efficient. The regression model had an LR chi2 value of 378.12, Pseudo R2 of 0.659, and *P* value < 0.0001, with a Brier score of 0.099 and tau-a value of 0.39. The AUC was 0.9146 (95 % CI 0.8879–0.9412), with a specificity of 92.7 %, sensitivity of 80.5 %, NPV of 85.21 %, and PPV of 90.17 % at the best cut-off value (Table 2).

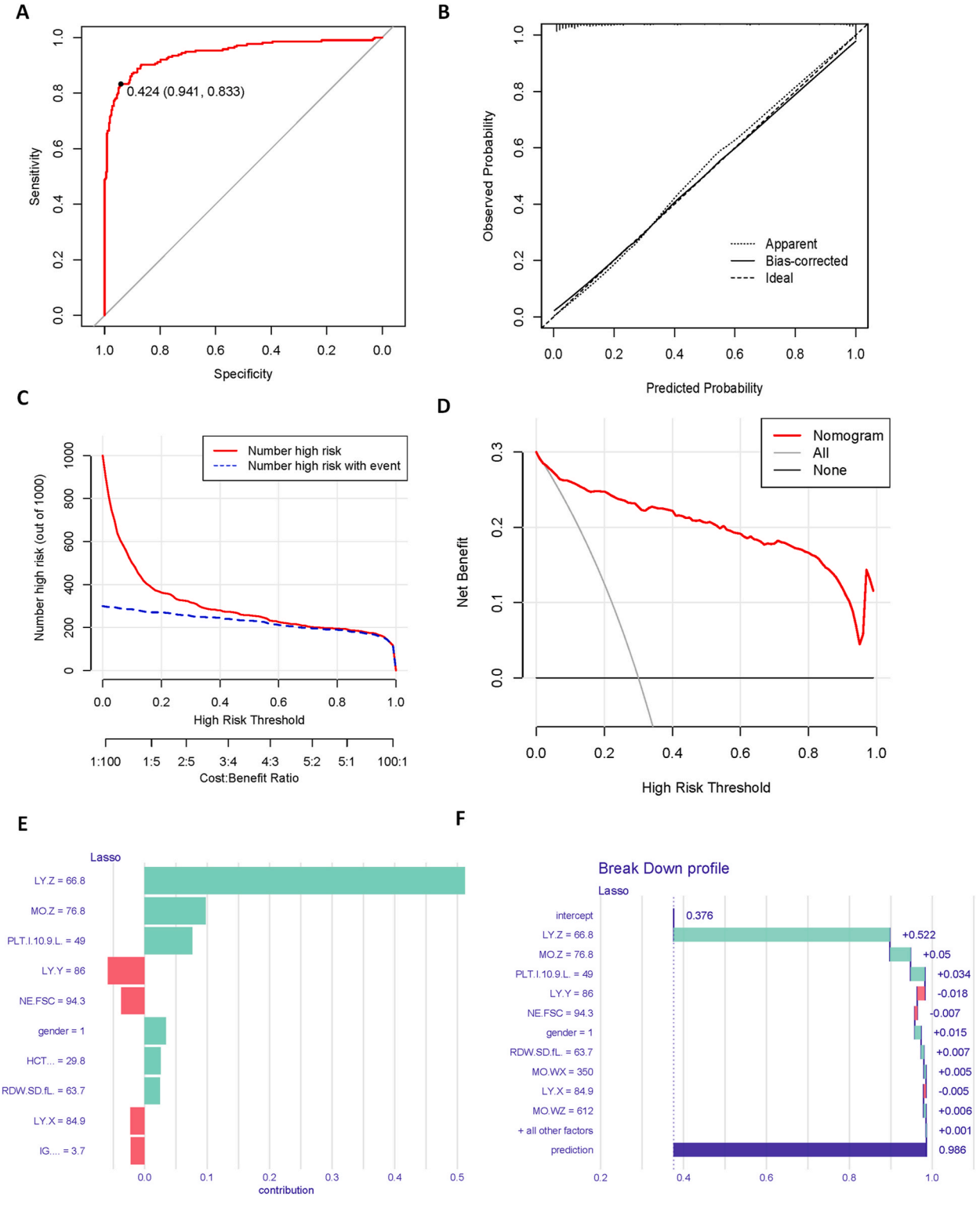
Following analysis by the Support Vector Machine algorithm with 10-fold cross-validation and recursive feature elimination (Fig. 2E), the top 20 optimal variables were ranked (Fig. 2F). The regression model showed an LR chi2 value of 385.09, a Pseudo R2 value of 0.668, and a *P* value < 0.0001. The AUC was 0.9132 (95%CI, 0.8851–0.9412). The specificity, sensitivity, NPV, and PPV by the optimal cut-off value were 95.00 %, 77.20 %, 86.73 %, and 90.56 %, respectively.

Based on the variable importance score ranking obtained from the XGBoost algorithm (FIG.SI), features such as LY-Z, MO-Z, and PLT-I were identified as key variables. The model's AUC, NPV, and PPV were 0.9182 (95%CI, 0.8914–0.9449), 86.18 %, and 91.95 % respectively.

After a comprehensive evaluation of the above models, it was found that LASSO ($\lambda = \min$) regression had the best differential diagnostic efficacy. It was further evaluated using independent validation data by comparing the predicted probabilities with clinically confirmed diagnoses. The AUC was found to be 0.9001 (95%CI: 0.8739–0.9263, $P < 0.001$) for distinguishing sepsis from local infection (Fig. 3A). When the best cut-off value was selected, the diagnostic specificity, sensitivity, NPV, and PPV of the model were found to be 86.70 %, 82.10 %, 84.04 %, and 79.46 % respectively. To further analyze the model, the calibration curve, clinical influence curve, and clinical decision curve of the model were plotted. The calibration curve results showed that the model's prediction probability was in good agreement with the actual probability (Fig. 3B). The clinical impact curve confirmed that when the risk threshold was more than 0.6, the predictive probability of the model was almost consistent with the actual occurrence probability of the disease (Fig. 3C). The clinical decision curve showed that when the threshold probability was more than 0.025, a better net benefit

Table 2
Diagnostic performance evaluation of model constructed by machine learning.

	LR	P	R2	Brier	tau-a	AUC	Specificity	Sensitivity	NPV	PPV
lasso1min	442.87	<0.01	0.734	0.08	0.418	0.9446 (0.9238–0.9654)	94.10 %	83.30 %	88.77 %	91.01 %
lasso1SE	414.35	<0.01	0.702	0.089	0.405	0.9306 (0.9069–0.9542)	89.60 %	84.70 %	86.67 %	89.56 %
RFIncMSE	376.89	<0.01	0.658	0.099	0.389	0.9138 (0.887–0.9405)	93.30 %	80.00 %	85.25 %	90.70 %
RFNode	378.12	<0.01	0.659	0.099	0.39	0.9146 (0.8879–0.9412)	92.70 %	80.50 %	85.21 %	90.17 %
SVM	385.09	<0.01	0.668	0.096	0.388	0.9132 (0.8851–0.9412)	95.00 %	77.20 %	86.73 %	90.56 %
XGBoost	389.65	<0.01	0.673	0.095	0.393	0.9182 (0.8914–0.9449)	86.80 %	86.50 %	86.18 %	91.95 %
Validation Cohort										
lasso1min	307.76	<0.01	0.577	0.122	0.373	0.9001 (0.8739–0.9263)	86.70 %	82.10 %	84.04 %	79.46 %
lasso1SE	308.16	<0.01	0.577	0.121	0.372	0.8986 (0.872–0.9252)	85.90 %	79.70 %	83.81 %	81.46 %
RFIncMSE	218.89	<0.01	0.441	0.15	0.312	0.8339 (0.7978–0.87)	80.50 %	72.90 %	78.16 %	78.52 %
RFNode	225.5	<0.01	0.452	0.148	0.313	0.8359 (0.8003–0.8714)	85.90 %	67.60 %	78.92 %	79.08 %
SVM	289.02	<0.01	0.55	0.128	0.357	0.883 (0.8536–0.9125)	84.70 %	78.30 %	81.38 %	79.29 %
XGBoost	158.23	<0.01	0.336	0.168	0.267	0.7859 (0.7442–0.8276)	72.90 %	73.40 %	75.81 %	80.31 %



(caption on next page)

Fig. 3. ROC curve, calibration curve, clinical impact curve, clinical decision curve, and local interpretation model of the best predictive model ($LASSO_{min}$). (3A): ROC curve, diagnostic specificity, and sensitivity under the best cutoff value. (3B): The calibration curve compares predicted and actual probabilities. The diagonal dotted line shows the ideal scenario. "Bias-corrected" refers to self-weightlifting sampling. (3C): Clinical impact curve: the x-axis represents the high-risk threshold, while the y-axis represents the number of risks in thousands of people. The predicted number of final events is displayed by the red curve of the model, while the actual number of risks is shown by the blue dotted line. (3D): The clinical decision curve is a graph that shows the benefits of various interventions. It measures the probability of success against the net benefit. The DCA curve indicates that the model is profitable when the threshold probability exceeds 0.025. (3E): Shaply additive interpretation (SHAP) calculates the average contribution of each variable by calculating all possible permutations between variables. (3F): Local interpretation's characteristic attribution analyzes differences between specific observations' predicted values and the model's average predicted values and how these differences are allocated to predictive variables. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

rate was achieved by the model (Fig. 3D).

3.3. Model interpretability

According to the SHAP analysis, the regression model constructed by LASSO indicated that LY-Z, MO-Z, and PLT-I had the most significant impact on individual outcomes (Fig. 3E). The Breakdown (BD) table revealed that the model's average predicted value for all samples was 0.376. Upon adding the LY-Z variable, the predicted value increased to 0.522, from the initial value of 0.375. Subsequently, adding MO-Z further increased the predicted value by 0.05 from the previous value, and this process continued until the final predicted probability for this sample reached 0.986. This approach facilitated the quantification of predictor variables' importance in both graphical and numerical terms, illustrating which variables had positive or negative effects (Fig. 3F).

4. Discussion

In this study, we constructed a prediction model for early sepsis diagnosis using multiple blood cell parameters. The model exhibited good differential diagnostic efficiency, with both the AUC_derivation and AUC_validation exceeding 0.9. This holds significant implications for assessing patients' conditions and specifying treatment strategies promptly. Meanwhile, the construction of a clinical prediction model based on multi-morphological parameters is reported for the first time, which to some extent deepens our understanding of the developmental patterns of blood cells in sepsis.

Several studies have highlighted complete blood count (CBC) as a valuable predictor of sepsis due to its rapidity and cost-effectiveness. Recently, Wong et al. utilized artificial neural networks based on clinical characteristics and CBC parameters to predict sepsis mortality in emergency departments [18]. Classical hematological analyzers were employed to determine cell volume, content (conductivity), and granularity (scatter; VCS) parameters [19,20]. Particularly, mean cell volumes of monocytes and neutrophils were identified as helpful parameters in distinguishing septic patients from healthy controls [21]. Furthermore, morphologic parameters could aid in sepsis management by monitoring treatment efficacy. Recent research suggests that MDW can serve as an early biomarker for predicting sepsis [22]. This exciting discovery inspires and guides us to focus on blood cell morphological changes in sepsis prediction and understand the underlying mechanisms. However, existing studies often concentrate on the diagnostic efficiency of a single indicator without conducting a comprehensive analysis of all morphological parameters. To address this issue, in addition to conventional CBC parameters such as white blood cell count, platelet count, hemoglobin content, etc., we have innovatively incorporated the characteristics of particles representing lymphocytes, monocytes, neutrophils, and other white blood cells, including nucleic acid content and cell volume size, into the analysis.

In recent years, machine learning has emerged as a novel tool in the medical field to analyze large amounts of data [23]. Shamim et al. constructed an interpretable machine learning model for accurate prediction of ICU sepsis [24]. Lei et al. utilized various machine learning algorithms, including LASSO, RF, and XGBoost, to predict the risk of death in patients with septic kidney injury [25]. Moreover, Jiang et al. developed predictive models for early prediction of sepsis-associated ARDS using LASSO, Light GBM, RF, and SVM [26].These clinical models, aided by algorithms, have demonstrated high predictive value. However, the reliance on traditional indicators like PCT and IL-6 is prevalent in most of these models. Unfortunately, not all patients can undergo the medical examinations required by these prediction models due to clinicians' subjective judgment and objective factors like patients' economic situations. This limitation hampers the generalization and application of these models. Daniel et al. developed a prediction model based on CBC parameters using an enhanced random forest algorithm, but overlooked the increasingly important CPD-related features [27]. In this study, we delve into CBC parameters, including cell population data, aiming to develop a simple and effective clinical prediction model for sepsis.

By integrating multiple machine learning methods, we have constructed a clinical prediction model that effectively distinguishes patients with sepsis from those with local infection. Among these methods, LASSO exhibited the best diagnostic performance (AUC = 0.9446). Utilizing LASSO (λ = min), we identified 24 candidate variables, which were further narrowed down to 10 variables through multiple logistic regression (gender, age, PLT-I, NE-FSC, LY-Y, LY-Z, MO-Y, MO-Z, NE-WZ, MO-WZ). According to the local interpretability model, the three most significant predictive factors were MO-Z, LY-Z, and PLT-I, associated with monocyte, lymphocyte, and platelet, respectively. Previous studies have suggested that biomarkers such as MDW and RDW can serve as valuable indicators for predicting sepsis. MDW, a parameter reflecting the variability and heterogeneity of monocyte size in peripheral blood, has been shown to assist in the early detection of sepsis patients in the emergency department. Interestingly, despite the different analytical

instruments used in this study, MO-Z, which reflects the heterogeneity of monocyte size, was also identified as a key variable. Furthermore, our interpretative algorithm indicates that LY-Z has a higher priority in the variable importance hierarchy than MO-Z, suggesting a potential importance of lymphocytes in biological processes and the development of diseases. When exploring the influence of LY-Z further, we have noticed its heterogeneity, which may reflect the functional status and diversity of responses within the lymphocyte population. This heterogeneity may indicate different levels of activation, immune reactions, and even prognostic markers for certain diseases. The discovery of greater heterogeneity in lymphocyte volume assigned a higher priority is not only novel but also potentially transformative. It opens up new avenues for research, allowing us to gain a deeper understanding of the mechanisms of lymphocyte function and dysfunction. Recently, related studies have demonstrated that T lymphocytes exhibit different morphological changes following exposure to various bacterial determinants, providing further support for our findings [28,29]. Additionally, platelets play a vital role in immune response, pathogen clearance, tissue repair, hemostasis, and thrombosis. Platelet-related parameters have proven useful in predicting the prognosis and mortality of sepsis [30].

Regarding CPD, the X-axis parameters related to internal complexity include NE-SSC, LY-X, MO-X, NE-WX, LY-WX, and MO-WX. The Y-axis indexes related to nucleic acid content include NE-SFL, LY-Y, MO-Y, NE-WY, LY-WY, and MO-WY. CPD provides quantitative information on the morphological and functional characteristics of neutrophils, monocytes, and lymphocytes, allowing for a detailed study of cellular morphological changes. Any pathophysiological changes that alter the morphology of leukocytes may influence CPD values, reflecting the morphological changes of cells in response to infection and inflammation. Moreover, a risk stratification scale, known as the neutrophils and monocytes (NEMO) scoring, based on CPD parameters, can be used to rapidly and reliably identify sepsis [13]. Furthermore, researchers have demonstrated that NE-SFL and MO-X significantly increased in patients with septic shock and exhibited good diagnostic efficacy (AUC, 0.75 and 0.72, respectively) [14]. we believe our study comprehensively utilizes CPD parameters to predict sepsis early.

However, this study still has several limitations. First, patients with autoimmune diseases, malignant tumors, viral hepatitis, and HIV infection were excluded due to potential immune dysfunction that could affect the characteristics of peripheral hemogram in these patients. Therefore, the prediction model is unsuitable for estimating the sepsis probability in these specific patients. Secondly, comorbidities, disease severity scores, and other clinical data were not fully accounted for, which limited our ability to conduct more detailed statistical analysis to a certain extent. Additionally, multi-center validation with larger sample sizes would increase the model's credibility.

5. Conclusions

Our team has successfully developed a sepsis prediction model using advanced machine learning algorithms. The model is based on 10 distinct CBC variables and has demonstrated a significant potential for early sepsis detection. This innovative approach has the potential to significantly improve patient outcomes by allowing for timely intervention and treatment. Further research is required to validate the model's efficacy, but the preliminary results are promising. Our findings represent a significant advancement in the field of sepsis detection and underscore the potential for machine learning to revolutionize healthcare.

Ethics declarations

All procedures performed in this study were by the Declaration of Helsinki and approved by the Ethics Committee of West China Hospital (No.2020-641).

Data availability statement

Data will be made available on request.

CRediT authorship contribution statement

Tiancong Zhang: Writing – original draft, Validation. **Shuang Wang:** Investigation. **Qiang Meng:** Investigation. **Liman Li:** Data curation. **Mengxue Yuan:** Methodology. **Shuo Guo:** Data curation. **Yang Fu:** Writing – original draft, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82002215) and the Natural Science Foundation of Sichuan Province (2023NSFSC1483). The funder was not involved in the design of the study, the collection, analysis, and interpretation of data, and in writing the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e34498.

References

[1] C. Caraballo, F. Jaimes, Organ dysfunction in sepsis: an Ominous trajectory from infection to death, Yale J. Biol. Med. 92 (4) (2019) 629–640.
[2] J.E. Gotts, M.A. Matthay, Sepsis: pathophysiology and clinical management, BMJ (Clinical research ed) 353 (2016) i1585, https://doi.org/10.1136/bmj.i1585.
[3] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, et al., The Third international Consensus Definitions for sepsis and septic shock (Sepsis-3), JAMA 315 (8) (2016) 801–810, https://doi.org/10.1001/jama.2016.0287.
[4] C. Pierrakos, D. Velissaris, M. Bisdorff, J.C. Marshall, J.L. Vincent, Biomarkers of sepsis: time for a reappraisal, Crit. Care 24 (1) (2020) 287, https://doi.org/10.1186/s13054-020-02993-5.
[5] E. Kyriazopoulou, L. Liaskou-Antoniou, G. Adamis, A. Panagaki, N. Melachroinopoulos, E. Drakou, et al., Procalcitonin to reduce long-term infection-associated adverse events in sepsis. A randomized trial, Am. J. Respir. Crit. Care Med. 203 (2) (2021) 202–210, https://doi.org/10.1164/rccm.202004-1201OC.
[6] A. Meghraoui-Kheddar, B.G. Chousterman, N. Guillou, S.M. Barone, S. Granjeaud, H. Vallet, et al., Two new neutrophil subsets define a discriminating sepsis signature, Am. J. Respir. Crit. Care Med. 205 (1) (2022) 46–59, https://doi.org/10.1164/rccm.202104-1027OC.
[7] Y. Lang, Y. Jiang, M. Gao, W. Wang, N. Wang, K. Wang, et al., Interleukin-1 receptor 2: a new biomarker for sepsis diagnosis and gram-negative/gram-positive bacterial differentiation, Shock 47 (1) (2017) 119–124, https://doi.org/10.1097/shk.0000000000000714.
[8] A.T. Cruz, P. Mahajan, B.K. Bonsu, J.E. Bennett, D.A. Levine, E.R. Alpern, et al., Accuracy of complete blood cell counts to identify febrile infants 60 Days or younger with invasive bacterial infections, JAMA Pediatr. 171 (11) (2017) e172927, https://doi.org/10.1001/jamapediatrics.2017.2927.
[9] C.P. Hornik, D.K. Benjamin, K.C. Becker, D.K. Benjamin Jr., J. Li, R.H. Clark, et al., Use of the complete blood cell count in late-onset neonatal sepsis, Pediatr. Infect. Dis. J. 31 (8) (2012) 803–807, https://doi.org/10.1097/INF.0b013e31825691e4.
[10] S. Zhang, X. Luan, W. Zhang, Z. Jin, Platelet-to-Lymphocyte and neutrophil-to-lymphocyte ratio as predictive biomarkers for early-onset neonatal sepsis, Journal of the College of Physicians and Surgeons–Pakistan : JCPSP 31 (7) (2021) 821–824, https://doi.org/10.29271/jcpsp.2021.07.821.
[11] L. Agnello, G. Bivona, M. Vidali, C. Scazzone, R.V. Giglio, G. Iacolino, et al., Monocyte distribution width (MDW) as a screening tool for sepsis in the Emergency Department, Clin. Chem. Lab. Med. 58 (11) (2020) 1951–1957, https://doi.org/10.1515/cclm-2020-0417.
[12] P. Hausfater, Boter N. Robert, C. Morales Indiano, M. Cancella de Abreu, A.M. Marin, J. Pernet, et al., Monocyte distribution width (MDW) performance as an early sepsis indicator in the emergency department: comparison with CRP and procalcitonin in a multicenter international European prospective study, Crit. Care 25 (1) (2021) 227, https://doi.org/10.1186/s13054-021-03622-5.
[13] E. Urrechaga, O. Bóveda, U. Aguirre, Role of leucocytes cell population data in the early detection of sepsis, J. Clin. Pathol. 71 (3) (2018) 259–266, https://doi.org/10.1136/jclinpath-2017-204524.
[14] P. Biban, M. Teggi, M. Gaffuri, P. Santuz, D. Onorato, G. Carpenè, et al., Cell population data (CPD) for early recognition of sepsis and septic shock in children: a pilot study, Frontiers in pediatrics. 9 (2021) 642377, https://doi.org/10.3389/fped.2021.642377.
[15] L. Freijeiro-González, M. Febrero-Bande, W. González-Manteiga, A critical review of LASSO and its derivatives for variable selection under dependence among covariates, Int. Stat. Rev. 90 (1) (2022) 118–145, https://doi.org/10.1111/insr.12469.
[16] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[17] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Their Appl. 13 (4) (1998) 18–28, https://doi.org/10.1109/5254.708428.
[18] B.P.K. Wong, R.P.K. Lam, C.Y.T. Ip, H.C. Chan, L. Zhao, M.C.K. Lau, et al., Applying artificial neural network in predicting sepsis mortality in the emergency department based on clinical features and complete blood count parameters, Sci. Rep. 13 (1) (2023) 21463, https://doi.org/10.1038/s41598-023-48797-9.
[19] D.H. Park, K. Park, J. Park, H.H. Park, H. Chae, J. Lim, et al., Screening of sepsis using leukocyte cell population data from the Coulter automatic blood cell analyzer DxH800, Int J Lab Hematol. 33 (4) (2011) 391–399, https://doi.org/10.1111/j.1751-553X.2011.01298.x.
[20] I.H. Celik, G. Demirel, H.T. Aksoy, O. Erdeve, E. Tuncer, Z. Biyikli, et al., Automated determination of neutrophil VCS parameters in diagnosis and treatment efficacy of neonatal sepsis, Pediatr. Res. 71 (1) (2012) 121–125, https://doi.org/10.1038/pr.2011.16.
[21] J. Mammen, J. Choudhuri, J. Paul, T.I. Sudarsan, T. Josephine, G. Mahasampath, et al., Cytomorphometric neutrophil and monocyte markers may strengthen the diagnosis of sepsis, J. Intensive Care Med. 33 (12) (2018) 656–662, https://doi.org/10.1177/0885066616682940.
[22] Y.H. Pan, H.W. Tsai, H.A. Lin, C.Y. Chen, C.C. Chao, S.F. Lin, et al., Early identification of sepsis-induced acute kidney injury by using monocyte distribution width, red-blood-cell distribution, and neutrophil-to-lymphocyte ratio, Diagnostics 14 (9) (2024), https://doi.org/10.3390/diagnostics14090918.
[23] A.L. Beam, I.S. Kohane, Big data and machine learning in health care, JAMA 319 (13) (2018) 1317–1318, https://doi.org/10.1001/jama.2017.18391.
[24] S. Nemati, A. Holder, F. Razmi, M.D. Stanley, G.D. Clifford, T.G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the ICU, Crit. Care Med. 46 (4) (2018) 547–553, https://doi.org/10.1097/ccm.0000000000002936.
[25] L. Dong, P. Liu, Z. Qi, J. Lin, M. Duan, Development and validation of a machine-learning model for predicting the risk of death in sepsis patients with acute kidney injury, Heliyon 10 (9) (2024) e29985, https://doi.org/10.1016/j.heliyon.2024.e29985.
[26] Z. Jiang, L. Liu, L. Du, S. Lv, F. Liang, Y. Luo, et al., Machine learning for the early prediction of acute respiratory distress syndrome (ARDS) in patients with sepsis in the ICU based on clinical data, Heliyon 10 (6) (2024) e28143, https://doi.org/10.1016/j.heliyon.2024.e28143.
[27] D. Steinbach, P.C. Ahrens, M. Schmidt, M. Federbusch, L. Heuft, C. Lübbert, et al., Applying machine learning to blood count data predicts sepsis with ICU admission, Clin. Chem. 70 (3) (2024) 506–515, https://doi.org/10.1093/clinchem/hvae001.
[28] K.L. Vom Werth, T. Wörmann, B. Kemper, P. Kümpers, S. Kampmeier, A. Mellmann, Investigating morphological changes of T-lymphocytes after exposure with bacterial determinants for early detection of septic conditions, Microorganisms 10 (2) (2022), https://doi.org/10.3390/microorganisms10020391.
[29] K.L. Vom Werth, B. Kemper, S. Kampmeier, A. Mellmann, Application of digital holographic microscopy to analyze changes in T-cell morphology in response to bacterial challenge, Cells 12 (5) (2023), https://doi.org/10.3390/cells12050762.
[30] C. Zhang, X. Shang, Y. Yuan, Y. Li, Platelet-related parameters as potential biomarkers for the prognosis of sepsis, Exp. Ther. Med. 25 (3) (2023) 133, https://doi.org/10.3892/etm.2023.11832.