Infection and Drug Resistance

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Development of a Predictive Model for Neonatal Hospital-Acquired Gastrointestinal Infections Utilizing Multiple Machine Learning Algorithms

Hui Shao[1], Huajuan Chen[2], Xiujuan Wang[1]

[1]Department of Infectology, Shaoxing Maternity and Child Health Care Hospital (Maternity and Child Health Care Affiliated Hospital, Shaoxing University), Shaoxing, People's Republic of China; [2]Department of Obstetrics, Shaoxing Maternity and Child Health Care Hospital (Maternity and Child Health Care Affiliated Hospital, Shaoxing University), Shaoxing, People's Republic of China

Correspondence: Xiujuan Wang, Department of Infectology, Shaoxing Maternity and Child Health Care Hospital (Maternity and Child Health Care Affiliated Hospital, Shaoxing University), Shaoxing, People's Republic of China, Email juliewang472@163.com

Objective: To develop and validate a predictive model for neonatal gastrointestinal infections using multiple machine learning algorithms.

Methods: We conducted a retrospective analysis of 176 neonates diagnosed with nosocomial gastrointestinal infections in NICU between 2020 and 2024, along with a randomly selected control group of 675 neonates without such diagnoses during their NICU stay. The study examined 29 perinatal and NICU treatment-related risk factors potentially associated with neonatal gastrointestinal infections. The dataset was randomly partitioned into training and testing sets. To address class imbalance and enhance minority class identification, we applied SMOTE to the training set.Feature selection used Boruta, Lasso, and Logistic regression, with consensus features from Venn analysis.Subsequently, eight machine learning algorithms were implemented to construct predictive models. Models were evaluated using AUC, F1 score, accuracy, sensitivity, and specificity.

Results: The model incorporated nine significant feature variables: gestational age, NE, PLT, central venous catheterization, nasogastric feeding, delivery mode, intrauterine distress, pregnancy-induced hypertension, and probiotic administration. Among the eight machine learning algorithms evaluated, the Neural Network model demonstrated optimal performance — achieving perfect metrics in the training set (AUC=0.895, F1=0.845, Accuracy=0.837, Sensitivity=0.888, Specificity=0.786, Precision=0.806) and robust results in the test set (AUC= 0.876, F1=0.862, Accuracy=0.856, Sensitivity=0.896, Specificity=0.817, Precision=0.830) — thus was selected as the final predictive model. Model interpretability was enhanced through SHAP analysis. Furthermore, a Shiny-based interactive web calculator for neonatal gastrointestinal infection risk prediction was successfully developed based on this model.

Conclusion: The model effectively identifies at-risk neonates early, supporting clinical decision-making and timely interventions.

Keywords: neonate, NICU, gastrointestinal infection, machine learning, neural network, shiny

Introduction

Neonatal Hospital-Acquired Gastrointestinal Infections (NHAGIs) are defined as gastrointestinal infections that develop in newborns more than 48 hours after hospital admission. These infections represent one of the most common and severe complications in neonatal intensive care unit (NICU), characterized by high incidence rates, significant mortality, and substantial treatment costs. Epidemiological studies indicate that NHAGIs account for 10–30% of all hospital-acquired infections in NICU, constituting a major contributor to neonatal mortality.[1] Regional studies in China have reported varying NHAGI incidence rates, with tertiary hospitals showing significantly lower rates compared to resource-limited settings, consistent with global disparities in neonatal care quality.[2] Globally, NHAGI burden varies widely, with higher rates in low-resource settings compared to high-income countries, underscoring the need for context-specific interventions.[3,4] The pathogenesis of NHAGIs is multifactorial, with key risk factors including preterm birth, low birth weight, invasive medical procedures, antibiotic usage and the inherent risk of cross-infection within the NICU

Received: 11 April 2025
Accepted: 7 August 2025
Published: 18 August 2025

environment.[5–7] Current predictive approaches predominantly depend on clinicians empirical judgments, which are limited by subjective interpretation, inadequate timeliness, and challenges in processing complex, high-dimensional data. These limitations significantly hinder effective early warning systems and precise clinical interventions.[8,9] Consequently, the development of an objective and reliable predictive tool has emerged as a critical priority in clinical research.

Machine learning (ML) possess the capability to autonomously extract patterns from extensive medical datasets and develop predictive models, offering significant advantages including enhanced objectivity, improved timeliness and robust processing capabilities for high-dimensional data.[10,11] The development of predictive models for NHAGIs utilizing ML enables early warning systems, facilitates personalized treatment strategies, and optimizes resource allocation. These capabilities contribute to reducing infection rates and mortality while improving treatment outcomes and enhancing the efficiency of medical resource utilization.[12,13] Specifically, Fleuren et al systematically demonstrated the utility of ML models (eg, random forests, gradient boosting) in sepsis prediction, achieving pooled AUCs of 0.82–0.88 across 28 studies, while Nemati et al developed an interpretable LSTM-based model for real-time ICU sepsis prediction with 85% sensitivity, highlighting ML's clinical translatability for infection-related outcomes.Recent research has demonstrated the application of various algorithms, including logistic regression, support vector machines, and random forests, in constructing predictive models for neonatal outcomes, with their efficacy being empirically validated.[14–16] These findings suggest that ML holds substantial promise for broad applications in the prediction and management of NHAGIs.

To address this research gap, we conducted a comprehensive investigation of multiple obstetric and NICU-related risk factors associated with NHAGIs. Utilizing feature selection techniques to identify significant risk factors, we developed and validated a machine learning-based predictive model for NHAGIs, demonstrating strong performance in predicting infection risk. This retrospective study was designed to elucidate critical obstetric and clinical diagnostic factors influencing NHAGIs occurrence. The findings from this research are expected to significantly contribute to clinical practice by providing valuable insights for early intervention strategies. Furthermore, this study offers substantial support for enhancing neonatal healthcare outcomes and promoting healthy infant development.

Methods

Study Population

The study population comprised 851 neonates admitted to the NICU of our hospital between January 2020 and December 2024. Among these, 176 neonates were diagnosed with NHAGIs during their NICU stay, while 675 neonates without NHAGIs diagnosis during the same period were randomly selected as controls. Patient inclusion criteria were established in accordance with the Diagnostic Guidelines for Nosocomial Infections issued by the National Health Commission of China.[17] Figure 1 presents the flow chart illustrating the study population screening process.

Data Collection

We retrospectively collected 29 perinatal and neonatal diagnostic and treatment-related characteristic variables from electronic medical records and nursing documentation systems.Missing data were handled using complete-case analysis, whereby any patient record with missing values for one or more variables was excluded from the analysis. This conservative approach was chosen to ensure the integrity of the analytical dataset.While we initially considered additional clinical factors such as antibiotic regimens (eg, specific agents, duration), detailed nutritional parameters (eg, caloric intake, growth velocity), and microbiome profiling, these variables were excluded due to either: (1) inconsistent documentation in medical records (>30% missing data), or (2) lack of standardized measurement protocols across cases.The included variables were:gestational age, birth weight, initial laboratory indicators upon NICU admission (WBC, HGB, PLT, NE), maternal age, maternal BMI, preeclampsia status, delivery mode, parity, timing of first feeding, milk source, nasogastric feeding, use of food additives, probiotic administration, intrauterine distress, hypothermia, hypoglycemia, asphyxia, respiratory failure, respiratory distress syndrome, ventilator use, central venous puncture, congenital heart defects, premature rupture of membranes, intrauterine infection, gestational diabetes, and pregnancy-

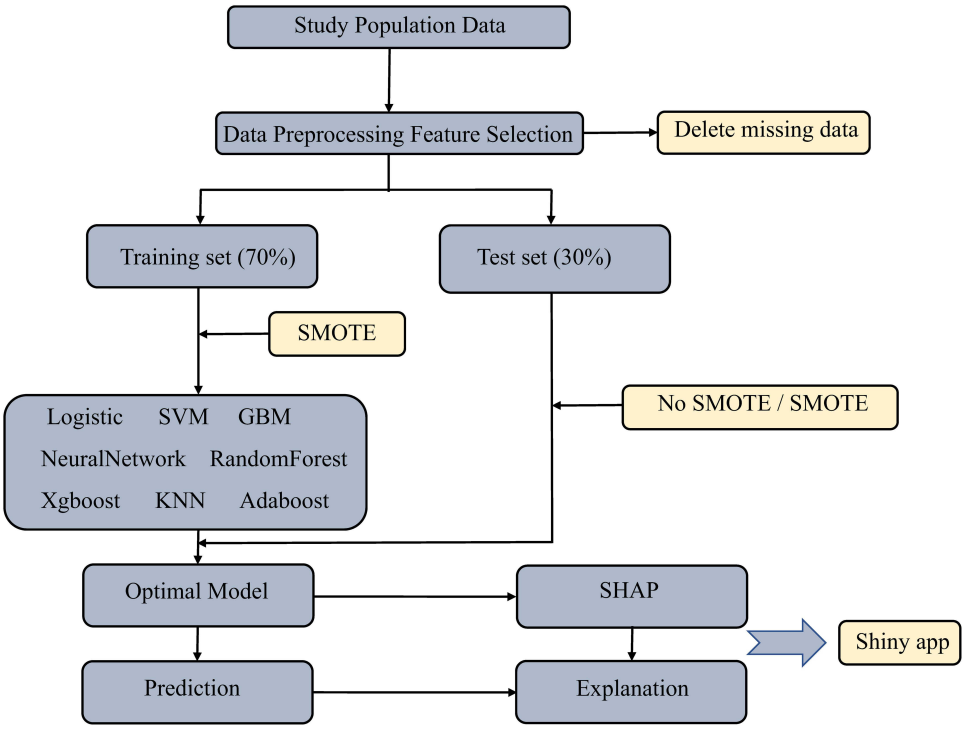

Figure 1 Flowchart of participant screening for study inclusion.

induced hypertension. These factors were comprehensively evaluated by a multidisciplinary team comprising experienced obstetricians, infectious disease specialists, and neonatologists, ensuring their clinical relevance and reference value.

Statistical Analysis

A comprehensive descriptive analysis was performed to characterize the study population. All statistical analyses were conducted using R-studio software. Continuous variables were expressed as mean ± standard deviation (SD) and analyzed using *t*-test, while categorical variables were expressed as percentages and analyzed using the chi-square test. Both univariate and multivariate logistic regression analyses were employed to calculate odds ratio (OR) with corresponding 95% confidence interval (CI), with statistical significance set at $p < 0.05$.The dataset was randomly partitioned into training and testing sets at a 7:3 ratio. Feature selection was performed in the training set using three distinct methods: Boruta, Lasso and logistic regression. The consensus features identified through these methods were determined using Venn diagram analysis, followed by correlation heatmap visualization to assess inter-variable relationships. Subsequently, eight machine learning algorithms were implemented in the training set, including Gradient Boosting Machine (GBM), Adaptive Boosting (Adaboost), Extreme Gradient Boosting (Xgboost), Neural Network, Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).For machine learning implementation, we employed a standardized training framework with 10-fold repeated (5 times) cross-validation (method="repeatedcv", number=10, repeats=5) and two-class summary metrics (classProbs=TRUE, summaryFunction=twoClassSummary). All models used the same random seed (set.seed(520)) for reproducibility. The eight algorithms were implemented with the following optimized hyperparameters: Logistic Regression: glm package with ridge regularization (alpha=0, lambda=0.01); SVM: svmRadial with sigma=0.01 and C=1 (optimized via grid search); GBM: n.trees=500, interaction.depth=3, shrinkage=0.01, n.minobsinnode=10; Neural Network: size=10, decay=0.01; Random Forest: mtry=sqrt(n_features), numRandomCuts=1; XGBoost: nrounds=500, max_depth=6, eta=0.01, gamma=0.1, colsample_bytree=0.8; KNN: kmax=20, distance=2; AdaBoost: mfinal=50, maxdepth=4.Model performance was evaluated and compared using multiple metrics: AUC-ROC, sensitivity, specificity, F1 score,

https://doi.org/10.2147/IDR.S533904

accuracyandprecision. The testing set was utilized for model validation and performance assessment. Based on its superior predictive performance, the neural network model was selected for the development of an online risk calculator using the Shiny framework. The calculator offers a user-friendly interface where clinicians can input variables and instantly receive NHAGIs risk predictions by clicking the "Predict" button.

Results

Patient's Characteristics and Logistic Regression Analyses

Table 1 presents the baseline characteristics of the training and validation cohorts. The patient data were randomly divided into training and validation sets at a 7:3 ratio. The training set (n = 946) underwent SMOTEprocessing, with Figure 2A illustrating the distribution of sample categories before and after SMOTE application. The validation set (n = 253) remained unprocessed for comparative analysis.Statistical analysis revealed significant differences ($p < 0.05$) in 12 characteristic

Table 1 Baseline Features of the Training and Test Set

Variables	Total (N = 1200)	Training Set (N = 946)	Test Set (N = 254)	P
Gestation week, Mean ± SD	35.88 ± 3.82	35.67 ± 3.82	36.68 ± 3.75	**<0.001**
Neonatal weight(g), Mean ± SD	2582.56 ± 892.96	2540.27 ± 892.33	2740.06 ± 879.25	**0.002**
WBC (10^9/L),Mean ± SD	13.28 ± 6.01	13.13 ± 6.07	13.86 ± 5.77	0.083
HGB(g/L),Mean ± SD	173.39 ± 19.03	173.24 ± 18.64	173.95 ± 20.46	0.599
PLT(10^9/L), Mean ± SD	257.00 ± 61.12	256.77 ± 61.07	257.87 ± 61.41	0.800
NE(%), Mean ± SD	50.87 ± 25.68	50.41 ± 27.88	52.55 ± 14.78	0.239
Maternal age,year, Mean ± SD	29.41 ± 4.49	29.49 ± 4.44	29.08 ± 4.67	0.196
BMI, Mean ± SD	27.24 ± 3.88	27.16 ± 3.93	27.51 ± 3.70	0.205
NHAGIs, n(%)				**<0.001**
No	675 (56.25)	473 (50.00)	202 (79.53)	
Yes	525 (43.75)	473 (50.00)	52 (20.47)	
Type of delivery, n(%)				**0.028**
Cesarean section	724 (60.33)	586 (61.95)	138 (54.33)	
Vaginal delivery	476 (39.67)	360 (38.05)	116 (45.67)	
Parity, n(%)				**0.002**
Single birth	931 (77.58)	716 (75.69)	215 (84.65)	
Twinbirth	269 (22.42)	230 (24.31)	39 (15.35)	
First feeding time, n(%)				0.149
Birthday	849 (70.75)	660 (69.77)	189 (74.41)	
Non-birthday	351 (29.25)	286 (30.23)	65 (25.59)	
Source of milk, n(%)				**0.010**
Maternal milk	239 (19.92)	198 (20.93)	41 (16.14)	
Mixture	746 (62.17)	594 (62.79)	152 (59.84)	
Formula milk	215 (17.92)	154 (16.28)	61 (24.02)	
Nasal feeding, n(%)				**<0.001**
No	872 (72.67)	664 (70.19)	208 (81.89)	
Yes	328 (27.33)	282 (29.81)	46 (18.11)	
Food additive, n(%)				0.120
No	1042 (86.83)	814 (86.05)	228 (89.76)	
Yes	158 (13.17)	132 (13.95)	26 (10.24)	
Add probiotics, n(%)				0.627
No	355 (29.58)	283 (29.92)	72 (28.35)	
Yes	845 (70.42)	663 (70.08)	182 (71.65)	
Intrauterine distress, n(%)				**0.005**
No	1089 (90.75)	870 (91.97)	219 (86.22)	
Yes	111 (9.25)	76 (8.03)	35 (13.78)	

(*Continued*)

Table 1 (Continued).

Variables	Total (N = 1200)	Training Set (N = 946)	Test Set (N = 254)	P
Hypothermia, n(%)				**0.039**
No	1176 (98.00)	923 (97.57)	253 (99.61)	
Yes	24 (2.00)	23 (2.43)	1 (0.39)	
Hypoglycemia, n(%)				0.351
No	933 (77.75)	741 (78.33)	192 (75.59)	
Yes	267 (22.25)	205 (21.67)	62 (24.41)	
Asphyxia, n(%)				0.424
No	1060 (88.33)	832 (87.95)	228 (89.76)	
Yes	140 (11.67)	114 (12.05)	26 (10.24)	
Respiratory failure, n(%)				**0.042**
No	860 (71.67)	665 (70.30)	195 (76.77)	
Yes	340 (28.33)	281 (29.70)	59 (23.23)	
Respiratory distress, n(%)				**0.006**
No	972 (81.00)	751 (79.39)	221 (87.01)	
Yes	228 (19.00)	195 (20.61)	33 (12.99)	
Ventilator use, n(%)				**0.005**
No	933 (77.75)	719 (76.00)	214 (84.25)	
Yes	267 (22.25)	227 (24.00)	40 (15.75)	
Central venous puncture, n(%)				0.332
No	1041 (86.75)	816 (86.26)	225 (88.58)	
Yes	159 (13.25)	130 (13.74)	29 (11.42)	
Congenital heart defect, n(%)				0.643
No	1109 (92.42)	876 (92.60)	233 (91.73)	
Yes	91 (7.58)	70 (7.40)	21 (8.27)	
Premature rupture of membranes, n(%)				0.443
No	870 (72.50)	681 (71.99)	189 (74.41)	
Yes	330 (27.50)	265 (28.01)	65 (25.59)	
Intrauterine infection, n(%)				0.084
No	1112 (92.67)	883 (93.34)	229 (90.16)	
Yes	88 (7.33)	63 (6.66)	25 (9.84)	
Gestational diabetes, n(%)				0.202
No	915 (76.25)	729 (77.06)	186 (73.23)	
Yes	285 (23.75)	217 (22.94)	68 (26.77)	
Hypertension during pregnancy, n(%)				0.233
No	1162 (96.83)	919 (97.15)	243 (95.67)	
Yes	38 (3.17)	27 (2.85)	11 (4.33)	
Pre-eclampsia, n(%)				0.577
No	1088 (90.67)	860 (90.91)	228 (89.76)	
Yes	112 (9.33)	86 (9.09)	26 (10.24)	

Notes: Bold text indicates statistical significance at P < 0.05; Continuous variables: mean ± SD (standard deviation); Categorical: n (%); N indicates the number of observations in each group; P values were derived from t-tests or χ^2-tests.

variables between the cohorts: gestational age, birth weight, NHAGIs status, delivery mode, parity, milk source, nasogastric feeding, intrauterine distress, hypothermia, respiratory failure, respiratory distress syndrome, and ventilator use.Table 2 displays the results of logistic regression analyses performed on the training set. The analysis identified several statistically significant predictive factors ($p < 0.05$), including gestational age, PLT, NE, delivery mode, milk source, nasogastric feeding, probiotic administration, intrauterine distress, central venous catheterizationand pregnancy-induced hypertension.

https://doi.org/10.2147/IDR.S533904

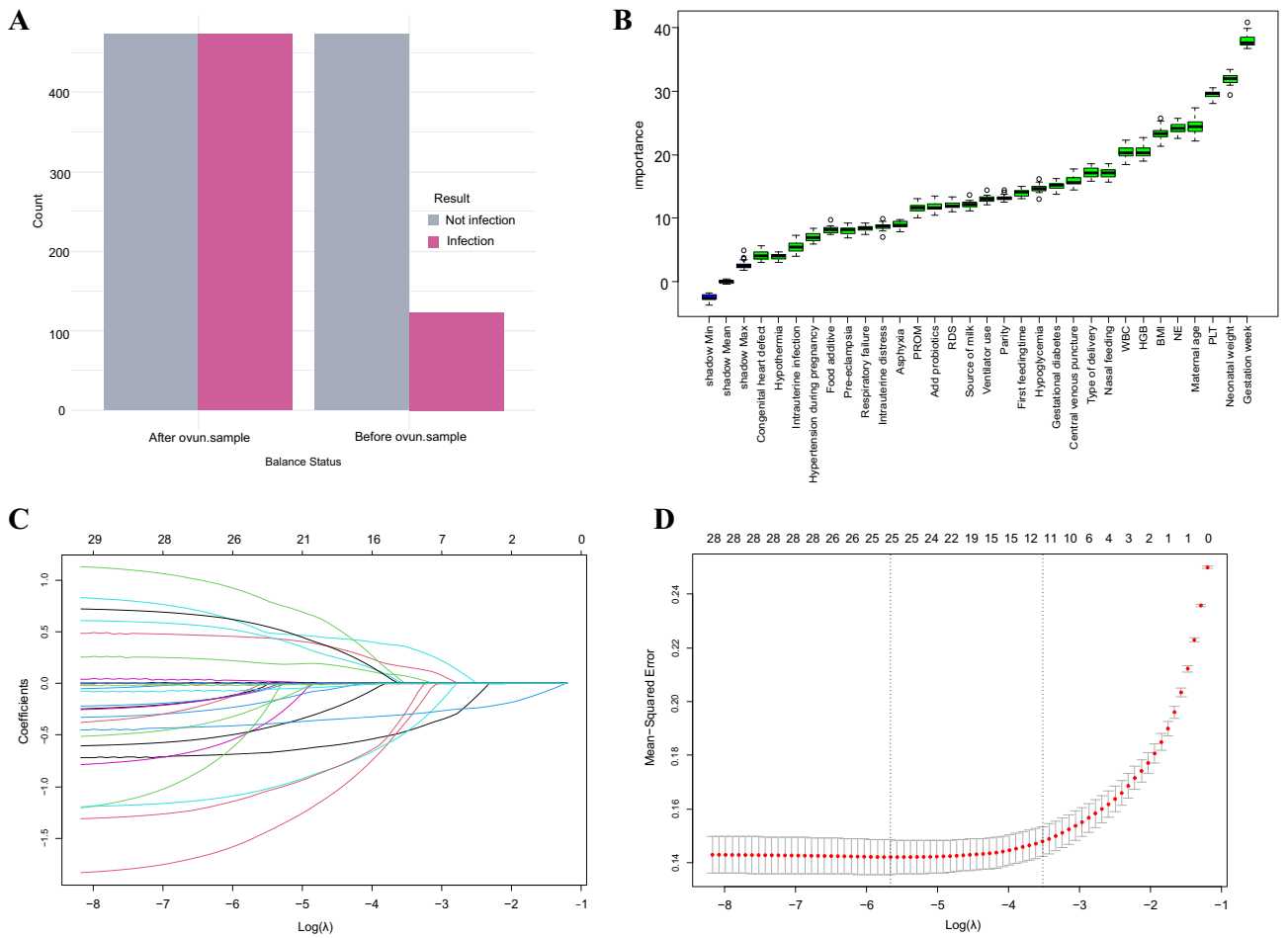

Figure 2 (**A**) The distribution of the class labels of the samples before and after applying SMOTE to the training set. (**B**) The result of Boruta algorithm screening important features. (**C** and **D**) The result of Lasso regression screening important features.

Broguta and Lasso Regression Analysis

The Boruta algorithm is an advanced feature selection method that effectively identifies significant predictors for modeling. In the Boruta visualization (Figure 2B), black box plots represent the minimum, average, and maximum Z-scores of shadow features, while green box plots denote confirmed features. The results demonstrate that all variables marked in green were identified as important features. Figure 2C presents the LASSO regression path diagram, displaying 29 distinct trajectories corresponding to the included variables. Each colored line represents the coefficient trajectory of an independent variable, with the y-axis indicating coefficient values and the lower x-axis showing log(λ) values. The upper x-axis displays the number of non-zero coefficients at each λ value. As log(λ) increases, the regression coefficients gradually converge toward zero. Figure 2D illustrates the cross-validation curve, where the x-axis represents log(λ) and the y-axis shows the binomial deviance. Through this process, LASSO regression identified 12 relevant variables, which were subsequently incorporated into the multivariate logistic regression analysis, as detailed in Table 3.

Venn Diagram and Spearman Correlation Heatmap

Through comparative analysis of feature selection results from Boruta, Lasso and logistic regression, we identified the intersecting subset of features common to all three methods as our final predictors. Figure 3A illustrates the nine feature variables ultimately selected for inclusion in the ML prediction model: gestational age, PLT, NE, delivery mode, nasogastric feeding, probiotic administration, intrauterine distress, central venous catheterization, and pregnancy-

Table 2 Univariate and Multivariate Logistic Regression Analysis of the Training Set

Variable	Univariate Logistic	P	Multivariable Logistic	P
	OR (95% CI)		OR (95% CI)	
Gestation week	0.61 (0.58–0.65)	**<0.001**	0.69 (0.62–0.78)	**<0.001**
Neonatal weight	1.00 (1.00–1.00)	**<0.001**	1.00 (1.00–1.00)	0.280
WBC	0.91 (0.89–0.94)	**<0.001**	1.04 (1.00–1.09)	0.070
HGB	1.00 (0.99–1.00)	0.199		
PLT	0.99 (0.99–0.99)	**<0.001**	0.99 (0.99–1.00)	**<0.001**
NE	0.95 (0.94–0.96)	**<0.001**	0.98 (0.97–1.00)	**0.028**
Maternal age	1.01 (0.98–1.04)	0.496		
BMI	0.99 (0.95–1.02)	0.400		
Type of delivery				
Cesarean section				
Vaginal delivery	0.28 (0.21–0.36)	**<0.001**	0.40 (0.25–0.62)	**<0.001**
Parity				
Single birth				
Twinbirth	4.89 (3.47–6.88)	**<0.001**	1.34 (0.83–2.16)	0.226
First feeding time				
Birthday				
Non-birthday	5.22 (3.81–7.16)	**<0.001**	1.25 (0.76–2.05)	0.374
Source of milk				
Mixture				
Formula milk	0.44 (0.30–0.65)	**<0.001**	0.47 (0.27–0.80)	**0.006**
Maternal milk	5.87 (3.92–8.78)	**<0.001**	0.76 (0.37–1.56)	0.451
Nasal feeding				
No				
Yes	12.66 (8.65–18.53)	**<0.001**	2.24 (1.07–4.71)	**0.032**
Food additive				
No				
Yes	7.50 (4.53–12.43)	**<0.001**	0.90 (0.36–2.25)	0.816
Add probiotics				
No				
Yes	1.58 (1.19–2.09)	**0.001**	0.61 (0.41–0.92)	**0.018**
Intrauterine distress				
No				
Yes	0.52 (0.32–0.85)	**0.009**	0.24 (0.11–0.55)	**<0.001**
Hypothermia				
No				
Yes	10.94 (2.55–46.91)	**0.001**	0.46 (0.06–3.65)	0.461
Hypoglycemia				
No				
Yes	1.30 (0.95–1.77)	0.098		
Asphyxia				
No				
Yes	3.03 (1.97–4.69)	**<0.001**	1.75 (0.92–3.36)	0.089
Respiratory failure				
No				
Yes	3.23 (2.40–4.35)	**<0.001**	0.83 (0.50–1.37)	0.460
Respiratory distress				
No				
Yes	11.82 (7.41–18.86)	**<0.001**	0.91 (0.37–2.23)	0.833

(*Continued*)

Table 2 (Continued).

Variable	Univariate Logistic	P	Multivariable Logistic	P
	OR (95% CI)		OR (95% CI)	
Ventilator use				
No				
Yes	10.54 (6.98–15.92)	**<0.001**	0.65 (0.26–1.63)	0.357
Central venous puncture				
No				
Yes	14.12 (7.50–26.58)	**<0.001**	2.69 (1.09–6.65)	**0.032**
Congenital heart defect				
No				
Yes	2.89 (1.68–4.97)	**<0.001**	0.72 (0.31–1.64)	0.435
Premature rupture of membranes				
No				
Yes	0.89 (0.67–1.18)	0.426		
Intrauterine infection				
No				
Yes	1.11 (0.66–1.85)	0.696		
Gestational diabetes				
No				
Yes	1.26 (0.93–1.70, p=0.142)	0.142		
Hypertension during pregnancy				
No				
Yes	0.34 (0.14–0.81)	**0.015**	0.14 (0.04–0.44)	**<0.001**
Pre-eclampsia				
No				
Yes	2.65 (1.63–4.31)	**<0.001**	0.81 (0.43–1.52)	0.504

Notes: Bold text indicates statistical significance at P < 0.05. P values from logistic regression; Multivariable model included variables with P<0.05 in univariate analysis.
Abbreviations: OR, odds ratio; CI, confidence interval.

Table 3 The Variables Selected by Multifactor Logistic Regression After the Lasso Regression

Variable	Estimate	Std.Error	OR (95% CI)	P
Type of delivery	−0.7131	0.2438	0.49 (0.30–0.79)	<0.05
Nasal feeding	0.6946	0.3466	2.00 (1.03–4.03)	<0.05
Add probiotics	−0.6031	0.2146	0.55 (0.36–0.83)	<0.05
Intrauterine distress	−1.7907	0.4536	0.17 (0.07–0.39)	<0.05
Central venous puncture	1.1059	0.4439	3.02 (1.29–7.46)	<0.05
Premature rupture of membranes	−1.2323	0.2565	0.29 (0.17–0.48)	<0.05
Gestational diabetes	0.7486	0.2326	2.11 (1.34–3.35)	<0.05
Hypertension during pregnancy	−1.2306	0.5889	0.29 (0.09–0.89)	<0.05
Gestation week	−0.4597	0.0714	0.63 (0.55–0.73)	<0.05
PLT	−0.0054	0.0018	0.99 (0.99–1.00)	<0.05
NE	−0.0188	0.0085	0.98 (0.97–1.00)	<0.05
BMI	−0.0809	0.0287	0.92 (0.87–0.97)	<0.05

Notes: P values from LASSO-selected multivariable logistic regression (all P<0.05).
Abbreviations: OR, odds ratio; CI, confidence interval.

induced hypertension. Figure 3B presents the Spearman correlation coefficient matrix heatmap, which was employed to assess inter-variable relationships.

Performance Comparison of Eight ML Algorithms

In the training set (Figure 3C), RF and KNN demonstrated perfect performance metrics (Accuracy = 1.000, Sensitivity = 1.000, Specificity = 1.000, Precision = 1.000, F1 = 1.000, AUC = 1.000), suggesting potential overfitting. GBM showed excellent performance with an AUC of 0.972 and an F1 score of 0.930, along with balanced metrics. XGBoost and Neural Network exhibited stable performance with AUCs of 0.905 and 0.892. SVM and AdaBoost demonstrated moderate performance, achieving AUCs of 0.882 and 0.864, while LR showed comparable performance with an AUC of 0.895.In the validation set

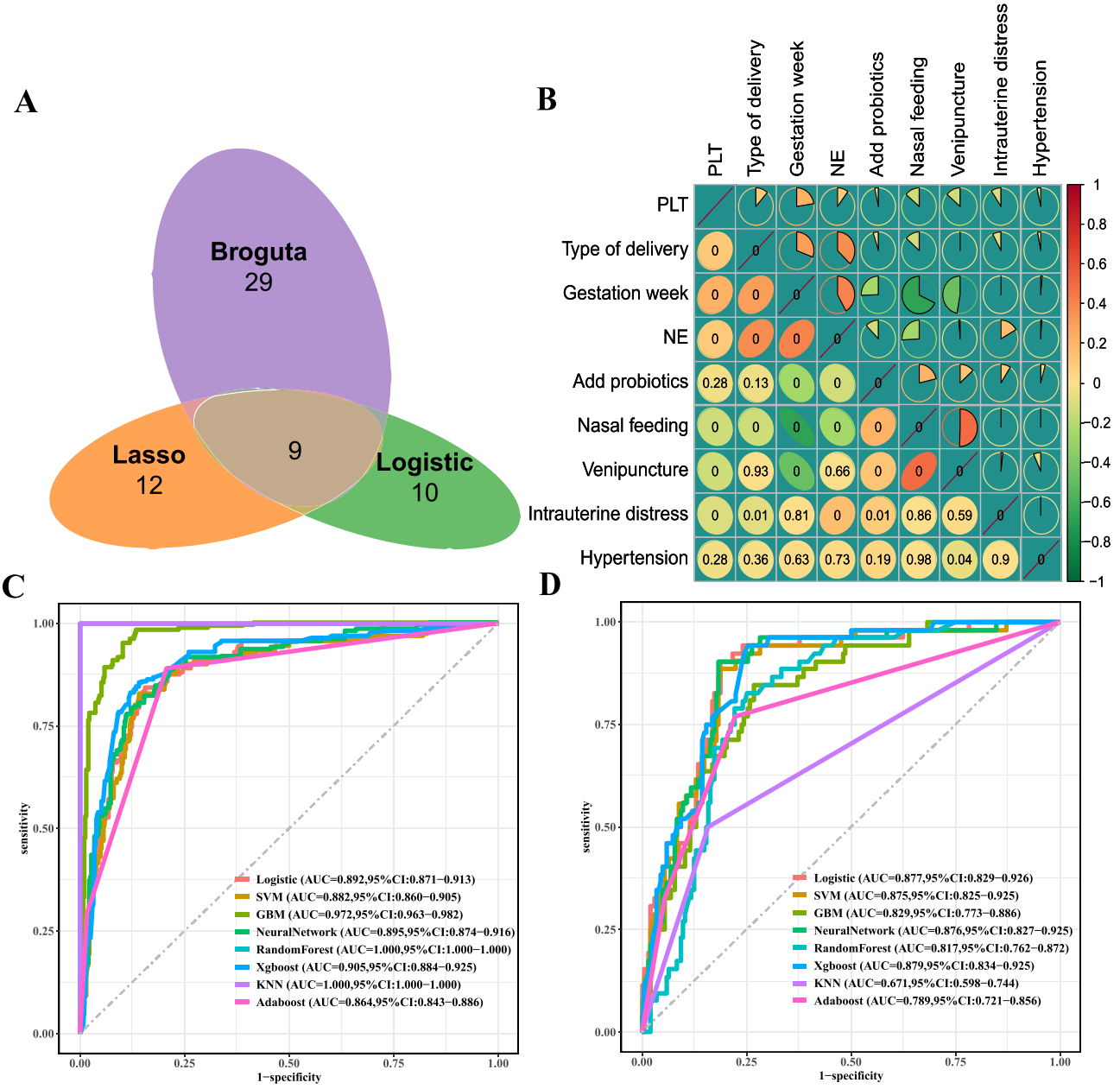

Figure 3 (**A**) Venn diagram of the three feature selection methods. (**B**) The Spearman correlation coefficient matrix heatmap. (**C** and **D**) ROC curves for eight models in the training cohort and validation cohort.

https://doi.org/10.2147/IDR.S533904

(Figure 3D), LR and Neural Network emerged as the top performers, achieving AUCs of 0.877 and 0.876, with balanced sensitivity and specificity. XGBoost attained an AUC of 0.790, demonstrating the highest sensitivity (0.942) but lower precision and F1, potentially indicating overfitting. SVM maintained stable performance with an AUC of 0.875, closely matching the top performers. GBM and RF showed reduced performance in the validation set, with AUCs of 0.829 and 0.817 respectively, likely due to overfitting. KNN exhibited the poorest generalization capability, achieving only an AUC of 0.671 and sensitivity of 0.500. AdaBoost demonstrated moderate performance with an AUC of 0.789, though its F1 score remained relatively low at 0.584.In summary, while the RF and KNN models demonstrated perfect performance on the training set, their significant performance degradation on the validation set indicated severe overfitting.The calibration curves for both training and validation sets revealed that the Neural Network model's predictions were closer to the ideal diagonal compared to other algorithms, indicating better alignment between predicted probabilities and actual outcomes (Figure 4A and B). DCA demonstrated that the RF model maintained high clinical net benefit across both datasets (Figure 4C and D). The classification performance of all ten algorithms was further evaluated through confusion matrix analysis (Figure 4E and F).Furthermore, ten-fold cross-validation performed on the training and validation sets demonstrated the Neural Network model's robust performance, with Figure 5A and B illustrating the AUC values for each fold and the average AUC. Similarly, GBM exhibited reduced AUC and F1 scores on the validation set, likely attributable to their high model complexity. In contrast, the Neural Network model maintained consistent performance across both training and validation sets, achieving high AUC and F1 scores, which supported its selection as the final predictive model.Notably, all models showed suboptimal F1 scores on the validation set, potentially due to the limited number of positive cases. This limitation was addressed through SMOTE oversampling of the validation set, which resulted in significant improvement of F1 scores (Figure 6).Based on these comprehensive evaluations, the Neural Network was ultimately selected as the optimal prediction model.

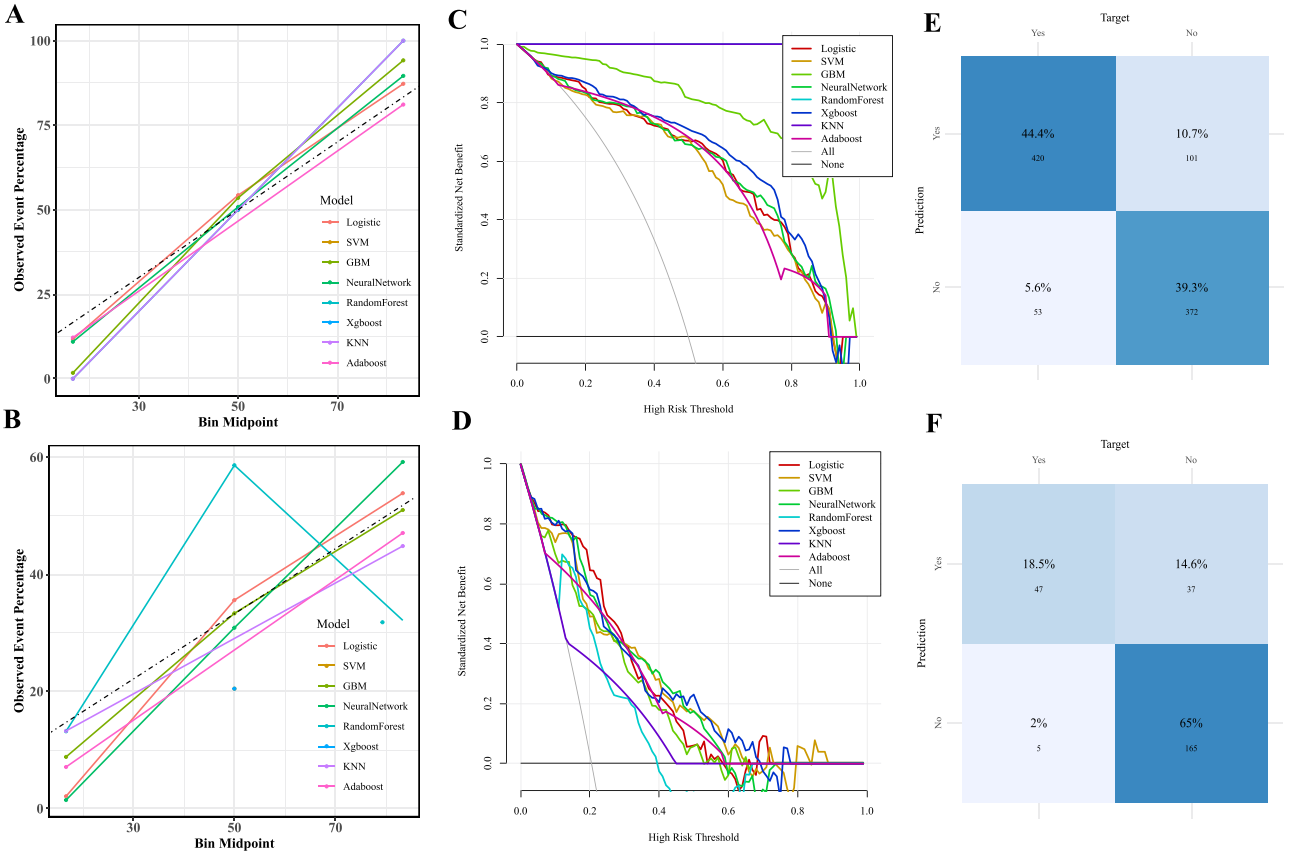

Figure 4 (**A** and **B**) Calibration curves for eight models in the training cohort and validation cohort. (**C** and **D**) DCA curves for eight models in the training cohort and validation cohort. (**E** and **F**) Confusion matrix results for eight models in the training cohort and validation cohort.

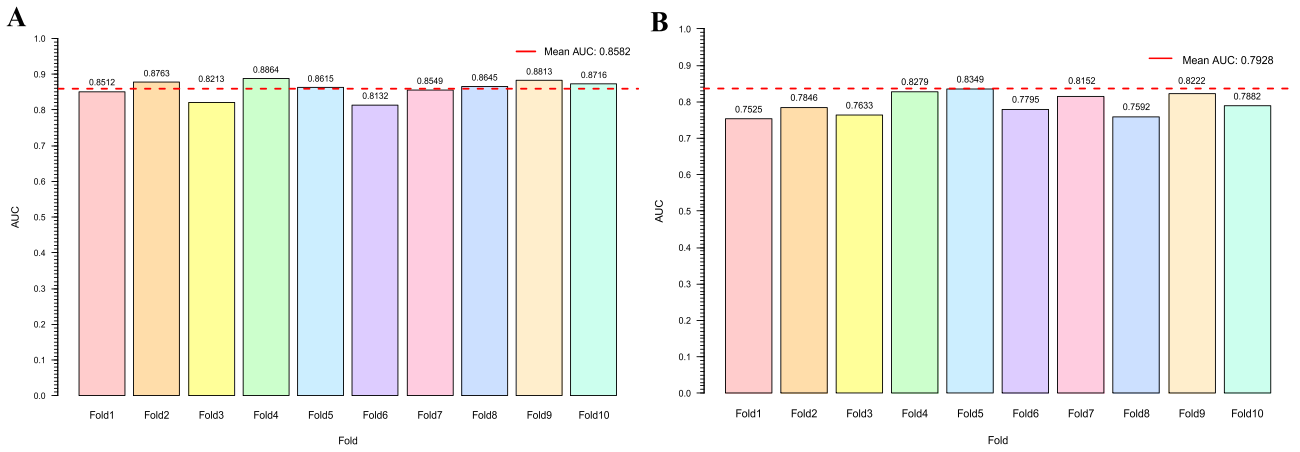

Figure 5 (**A**) Ten-fold cross-validation of the training cohort. (**B**) Ten-fold cross-validation of the validation cohort.

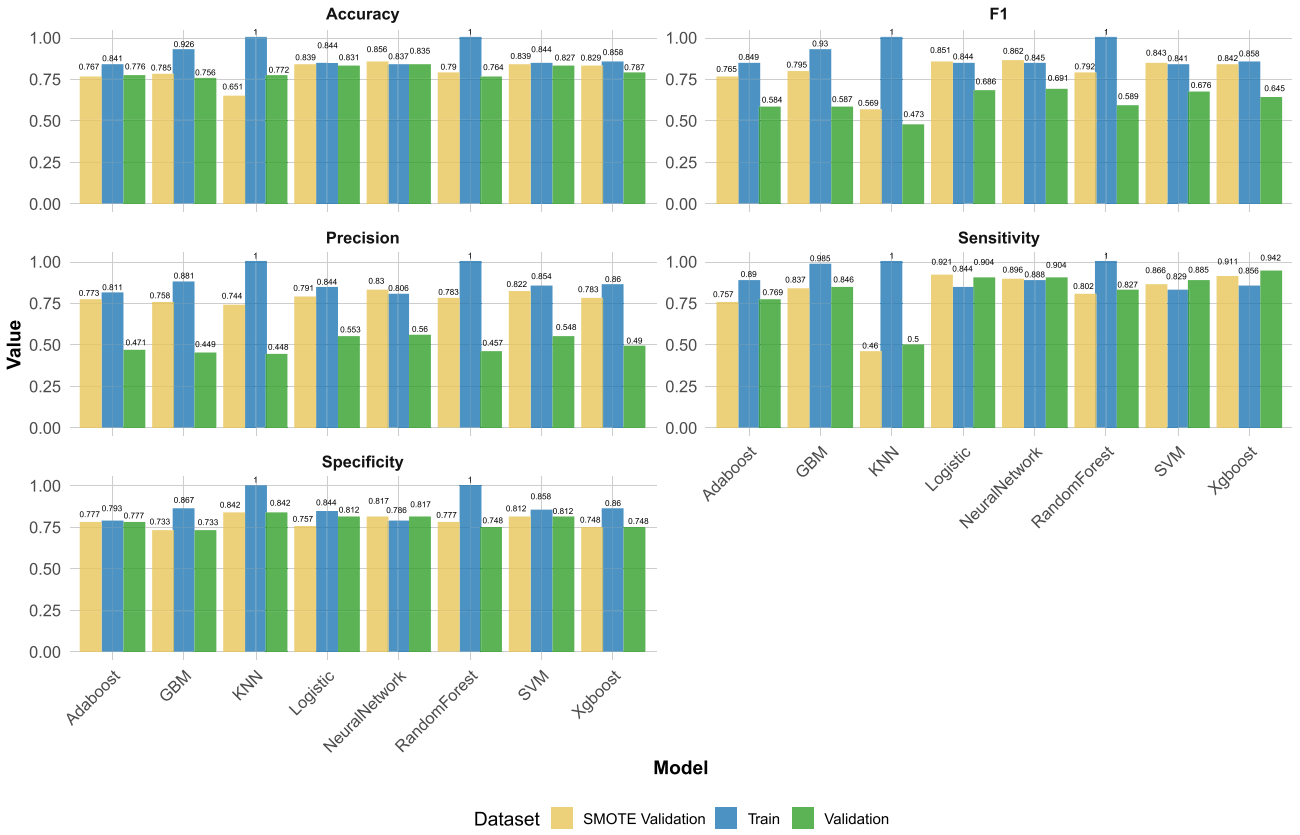

Figure 6 The performance metrics of the eight machine learning models across the training, validation, and SMOTE-processed validation sets.

Characteristic Importance and Interpretability of Model

Figure 7A presents the feature importance scores of the Neural Network model, visualized using SHapley Additive exPlanations (SHAP) values. The marginal contributions of the nine selected features are illustrated across all samples in Figure 7B. The analysis reveals that gestational age, PLT, NE, delivery mode, and nasogastric feeding demonstrate dispersed sample distributions and wider SHAP value ranges, indicating their substantial impact on model predictions. Conversely, intrauterine distress, central venous catheterization, and pregnancy-induced hypertension show distributions

https://doi.org/10.2147/IDR.S533904

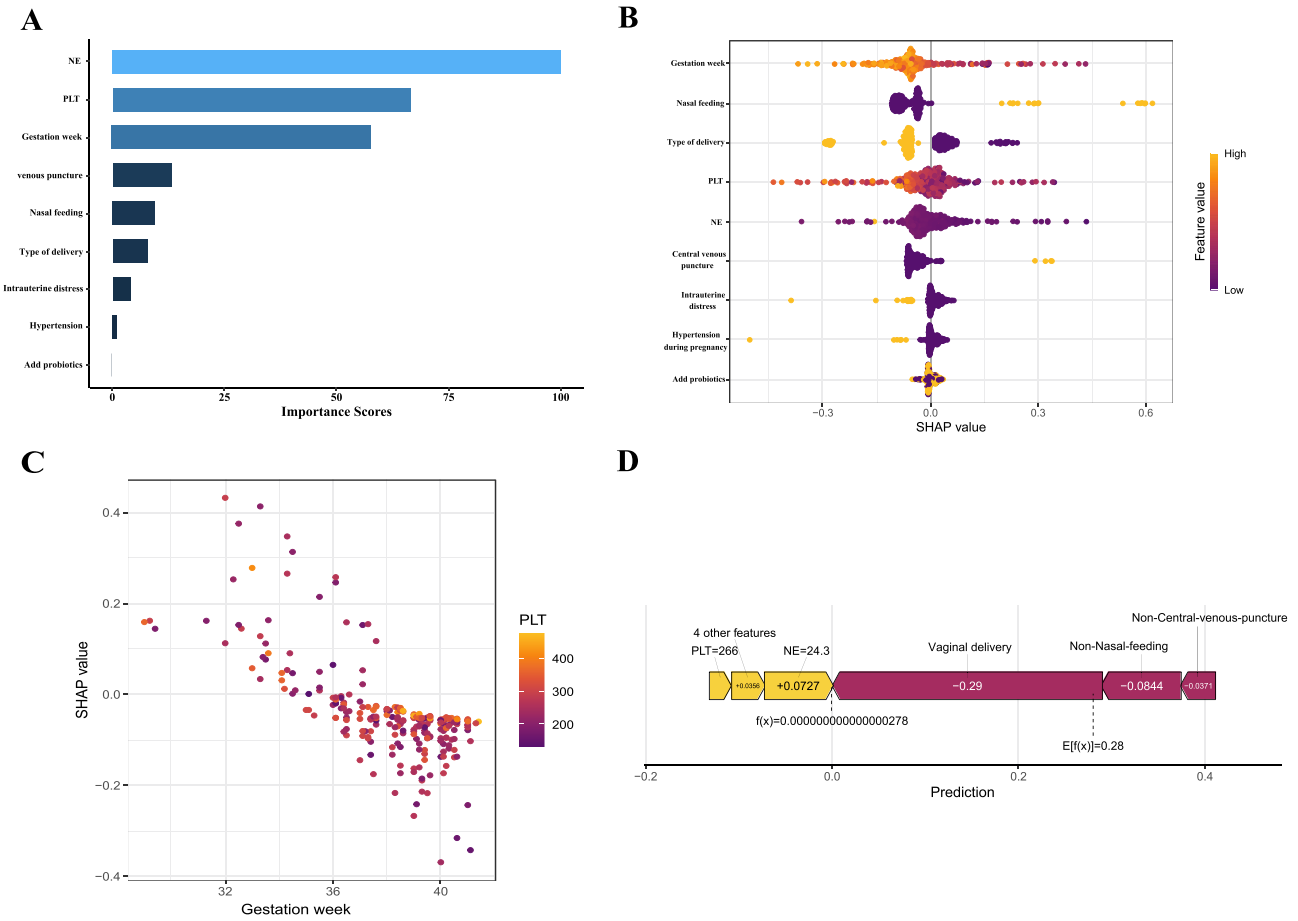

Figure 7 (**A**) Ranking of the importance scores for the eight variables. (**B**) Swarm map based on SHAP interpretation. (**C**) Dependence plot between the characteristic variables gestation week and PLT. (**D**) Single-sample interpretable force diagram.

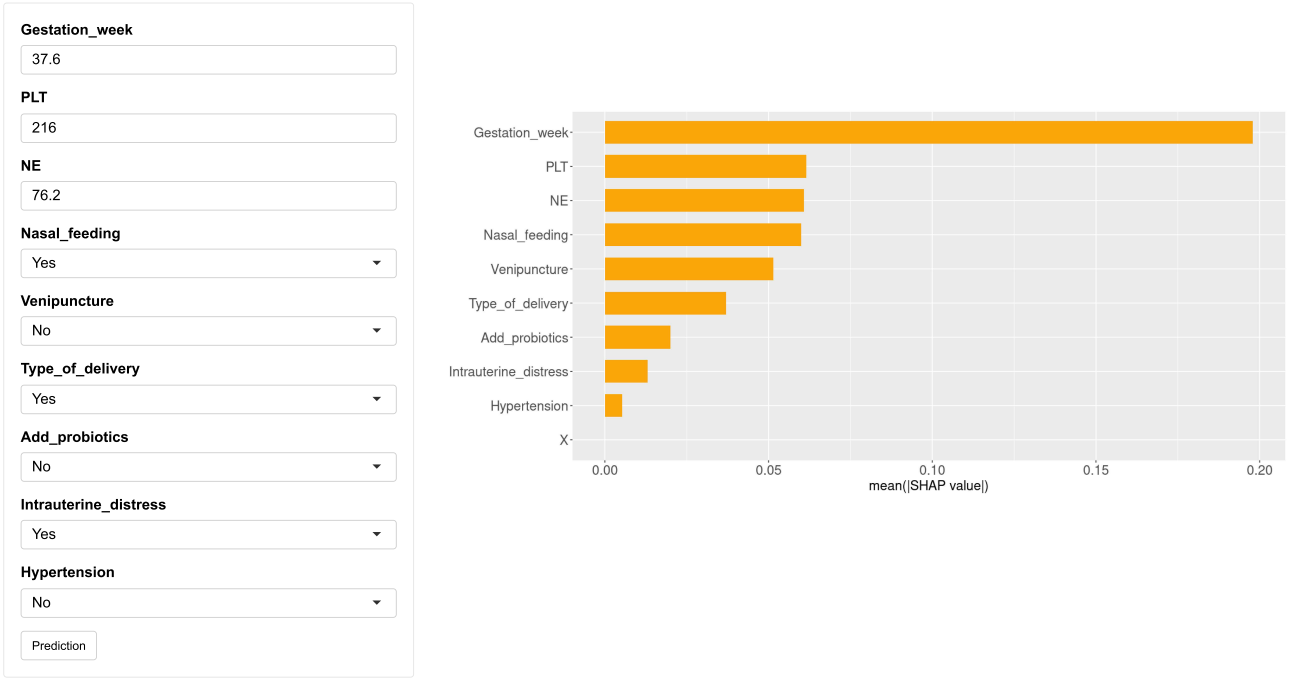

Figure 8 User interface for a risk calculator developed using shiny.

concentrated near SHAP = 0, suggesting minimal influence on the model's output.The dependency plot (Figure 7C) demonstrates that platelet count (particularly within the range of 300–400 × 10^9/L) may significantly contribute to prediction outcomes at specific gestational ages (eg, 36 weeks). Additionally, a force plot was generated to provide detailed feature-level explanations for individual sample predictions, as shown in Figure 7D.

Shiny Application of the Model

We developed an interactive risk calculator using Shiny in the R programming language to facilitate clinical implementation of the NHAGIs prediction model. This web-based application, accessible at [http://sh15609631795.shinyapps.io/EnteritisPredictor/], provides a user-friendly interface for clinical decision support. Figure 8 demonstrates the application interface, where clinicians can input relevant patient data and obtain the probability of NHAGIs by clicking the "Predict" button.

Discussion

Neonatal hospital-acquired gastrointestinal infections (NHAGIs) result from the complex interplay of multiple risk factors. In this study, we developed and validated eight ML algorithms to predict NHAGIs using comprehensive data encompassing various diagnostic and clinical factors. Among these models, the Neural Network demonstrated consistently superior performance across both training and testing datasets. This model's effectiveness was further supported by DCA and calibration curve results, indicating strong clinical applicability.The clinical significance of our ML approach lies in its ability to identify critical risk factors associated with NHAGIs. Our analysis identified nine key predictive factors: gestational age, NE, PLT, central venous catheterization, nasogastric feeding, delivery mode, intrauterine distress, pregnancy-induced hypertension and probiotic administration.The clinical significance of our ML approach is underscored by its superior performance compared to existing non-ML predictive tools. Traditional clinical scoring systems like the Modified Neonatal Sepsis Score (MNSS) demonstrate limited predictive accuracy due to their reliance on 6–8 predefined variables through logistic regression.[18,19] In contrast, our neural network model achieved significantly higher discrimination while incorporating 29 input features, capturing complex interactions missed by conventional approaches. Notably, our model identified several novel predictors absent in current scores,[20] and overcame the classic accuracy-interpretability trade-off through SHAP analysis. The development of our Shiny application further addresses a critical implementation gap, enabling real-time risk assessment unavailable in manual scoring systems.[21] These advancements suggest ML can overcome key limitations of rule-based tools while maintaining clinical interpretability.

Based on the permutation importance analysis of feature variables in the neural network model, gestational age, nasogastric feeding, and delivery mode emerged as crucial predictors of neonatal nosocomial gastrointestinal infections. Gestational age serves as a fundamental indicator of neonatal health, with prematurity being strongly associated with increased infection risk. Preterm infants often exhibit immature immune systems and compromised intestinal barrier function, rendering them more susceptible to gastrointestinal infections.[22] Previous research has established a correlation between nasogastric feeding and neonatal gastrointestinal infections. The use of nasogastric tubes may facilitate pathogen entry into the gastrointestinal tract, particularly with prolonged usage.[23,24] Furthermore, cesarean delivery has been associated with elevated infection risk compared to vaginal delivery, potentially due to its impact on the establishment of neonatal gut microbiota, thereby increasing vulnerability to gastrointestinal infections.[25,26] Hematological indicators, NE and PLT counts, also play essential roles in predicting NHAGIs. Neutrophils, as critical components of the immune system, exhibit count variations that reflect infection or inflammatory status. The degree of neutrophil elevation in infected neonates specifically correlates with infection severity and type.[27] Similarly, platelet counts serve as indicators of immune status and inflammatory response. Notably, thrombocytopenia frequently occurs in neonates, particularly during infections or inflammatory states, potentially exacerbating infection risk.[28] Our findings also indicate that probiotic supplementation in breast milk or formula may inhibit the development of gastrointestinal infections in newborns by modulating intestinal microbiota. Previous research has demonstrated that probiotics can strengthen intestinal barrier function and suppress pathogenic growth.[29] Similarly, central venous catheterization, like nasogastric tube placement, represents a significant risk factor for nosocomial infections in neonates, as these devices can serve as entry points for pathogens.[30] Intrauterine distress has been shown to potentially compromise neonatal

immune function through hypoxia, increasing infection susceptibility. The resultant intrauterine hypoxia may also impair intestinal function, further elevating the risk of gastrointestinal infections.[31] Notably, our analysis revealed an unexpected association between pregnancy-induced hypertension (PIH) and reduced NHAGIs risk, which may be attributed to several factors:First, PIH (including preeclampsia) often necessitates medically indicated preterm delivery, resulting in more intensive medical supervision and prophylactic antibiotic use for preterm infants, potentially lowering gastrointestinal infection risk.[32] The stringent infection control measures implemented in NICU, including aseptic techniques and antibiotic prophylaxis, may further contribute to this protective effect.[33] Second, PIH may influence neonatal immunity through placental transmission of immune-modulating factors, potentially enhancing immune defense mechanisms.[34] Third, the increased frequency of prenatal monitoring and interventions in PIH cases may indirectly reduce neonatal infection risk by enabling earlier detection and management of intrauterine infections or other complications.[35] Finally, mothers with PIH may be more likely to initiate breastfeeding, which has been consistently associated with reduced NHAGIs risk. Breast milk contains immunoglobulins and beneficial microorganisms that support the development of neonatal intestinal barrier function.[36]

This study developed and evaluated eight ML algorithms to create a predictive model for NHAGIs using comprehensive hospital data. We systematically analyzed the relationship between NHAGIs and various obstetric and clinical parameters, while employing advanced techniques to interpret the model's decision-making process and address potential issues of model interpretability. However, several limitations should be acknowledged. First, the model's generalizability may be limited as it was developed using data from a single center. Second, despite achieving an accuracy exceeding 85%, the model requires validation through prospective studies to further establish its clinical utility and practical applicability.Third, while critical factors like antibiotic regimens and microbiome data were considered, they were excluded from the final model due to either inconsistent documentation in medical records (>30% missing data) or lack of standardized measurement protocols, which may represent potential omitted variable bias that should be addressed in future studies with more comprehensive data collection protocols.

Conclusion

In conclusion, our Neural Network model effectively predicts neonatal hospital-acquired gastrointestinal infections, identifying critical risk factors and enabling early intervention via a Shiny-based tool. This advances neonatal care by supporting timely clinical decisions. Prospective, multi-center validation is needed to confirm generalizability and optimize clinical integration, addressing limitations and enhancing the model's impact on reducing NHAGI-related morbidity.

Abbreviations

NHAGIs,neonatalhospital-acquired gastrointestinal infections;ML,machinelearning;SHAP,shapleyadditive explanations; NICU, neonatal intensive care unit;SMOTE,syntheticminorityover-sampling technique; AUC, Area under the curve; CI, confidence interval; WBC, white blood cell; HCG, Hemoglobin;PLT, platelet; NE, neutrophil;BMI,bodymassindex; PROM, Premature rupture of membranes.

Data Sharing Statement

Relevant data from this study can be obtained from the corresponding author.

Ethics Statement

This retrospective study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Shaoxing Maternal and Child Health Hospital (IRB-AF-022-01.5). The research involved analysis of anonymized medical records and did not include human participants or animal trials. The ethics committee waived the requirement for informed consent given the retrospective nature of the study. All data were handled in compliance with institutional guidelines and regulations for patient data confidentiality.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

There is no supporting funding.

Disclosure

No conflict of interest is declared.

References

1. Stoll BJ, Hansen NI, Bell EF, et al. Trends in care practices, morbidity, and mortality of extremely preterm neonates, 1993–2012. *JAMA*. 2015;314(10):1039–1051. PMID: 26348753. doi:10.1001/jama.2015.10244
2. Zaidi AK, Huskins WC, Thave D, et al. Hospital-acquired neonatal infections in developing countries. *Lancet*. 2005;365(9465):1175–1188. PMID: 15794973. doi:10.1016/S0140-6736(05)71881-X
3. Allegranzi B, Bagheri Nejad S, Combescure C, et al. Burden of endemic health-care-associated infection in developing countries: systematic review and meta-analysis. *Lancet*. 2011;377(9761):228–241. PMID: 21146207. doi:10.1016/S0140-6736(10)61458-4
4. Rosenthal VD, Al-Abdely HM, El-Kholy AA, et al. International nosocomial infection control consortium report, data summary of 50 countries for 2010–2015: device-associated module. *AM J Infect Control*. 2016;44(12):1495–1504. PMID: 27742143. doi:10.1016/j.ajic.2016.08.007
5. Boghossian NS, Page GP, Bell EF, et al. Late-onset sepsis in very low birth weight infants from singleton and multiple-gestation births. *J Pediatr-US*. 2013;162(6):1120–1124. PMID: 23324523. doi:10.1016/j.jpeds.2012.11.089
6. Patel RM, Knezevic A, Shenvi N, et al. Association of red blood cell transfusion, anemia, and necrotizing enterocolitis in very low-birth-weight infants. *JAMA*. 2016;315(9):889–897. PMID: 26934258. doi:10.1001/jama.2016.1204
7. Cotten CM, Taylor S, Stoll B, et al. Prolonged duration of initial empirical antibiotic treatment is associated with increased rates of necrotizing enterocolitis and death for extremely low birth weight infants. *Pediatrics*. 2009;123(1):58–66. PMID: 19117. doi:10.1542/peds.2007-3423
8. Saria S, Butte A, Sheikh A. Better medicine through machine learning: what's real, and what's artificial? *PLOS Med*. 2018;15(12):e1002721. PMID: 30596635. doi:10.1371/journal.pmed.1002721
9. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2018;19(6):1236–1246. PMID: 28481991. doi:10.1093/bib/bbx044
10. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24–29. PMID: 30617335. doi:10.1038/s41591-018-0316-z
11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56. PMID: 30617339. doi:10.1038/s41591-018-0300-7
12. 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. *Intens Care Med*. 2020;46(3):383–400. PMID: 31965266. doi:10.1007/s00134-019-05872-y
13. Nemati S, Holder A, Razmi F, et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46(4):547–553. PMID: 29286945. doi:10.1097/CCM.0000000000002936
14. Xu H, Peng X, Peng Z, et al. Construction and SHAP interpretability analysis of a risk prediction model for feeding intolerance in preterm newborns based on machine learning. *BMC Med Inform DecisMak*. 2024;24(1):342. PMID: 39558307. doi:10.1186/s12911-024-02751-5
15. Tan X, Zhang X, Chai J, et al. Constructing a predictive model for early-onset sepsis in neonatal intensive care unit newborns based on SHapley Additive exPlanations explainable machine learning. *TranslPediatr*. 2024;13(11):1933–1946. PMID: 39649648. doi:10.21037/tp-24-278
16. Wang D, Huang S, Cao J, et al. A comprehensive study on machine learning models combining with oversampling for bronchopulmonary dysplasia-associated pulmonary hypertension in very preterm infants. *Respir Res*. 2024;25(1):199. PMID: 38720331. doi:10.1186/s12931-024-02797-z
17. Chinese Society of Surgical Infection and Intensive Care, Chinese Society of Surgery, Chinese Medical Association; Chinese College of Gastrointestinal Fistula Surgeons, Chinese College of Surgeons, Chinese Medical Doctor Association. [Chinese guideline for the prevention of surgical site infection]. *Zhonghua Wei Chang Wai Ke Za Zhi*. 2019;22(4):301–314 . doi:10.3760/cma.j.issn.1671-0274.2019.04.001
18. Escobar GJ, Puopolo KM, Wi S, et al. Stratification of risk of early-onset sepsis in newborns ≥ 34 weeks' gestation. *Pediatrics*. 2014;133(1):30–36. PMID: 24366992. doi:10.1542/peds.2013-1689
19. Wynn JL, Wong HR, Shanley TP, et al. Time for a neonatal-specific consensus definition for sepsis. *Pediatr Crit Care Me*. 2014;15(6):523–528. PMID: 24751791. doi:10.1097/PCC.0000000000000157
20. Zea-Vera A, Ochoa TJ. Challenges in the diagnosis and management of neonatal sepsis. *J Trop Pediatrics*. 2015;61(1):1–13. PMID: 25604489. doi:10.1093/tropej/fmu079
21. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25(9):1337–1340. PMID: 31427808. doi:10.1038/s41591-019-0548-6
22. Chawanpaiboon S, Vogel JP, Moller AB, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health*. 2019;7(1):e37–e46. PMID: 30389451. doi:10.1016/S2214-109X(18)30451-0

23. Sitomorang I, Garna H. Risk factors for nosocomial infections in neonatal intensive care unit. *Paediatr Indones*. 2018;34(1–2):48–56. doi:10.14238/pi34.1-2.1994.48-56
24. Auriti C, Maccallini A, Di Liso G, Di Ciommo V, Ronchetti MP, Orzalesi M. Risk factors for nosocomial infections in a neonatal intensive-care unit. *J Hosp Infect*. 2003;53(1):25–30. PMID: 12495682. doi:10.1053/jhin.2002.1341
25. Shao Y, Forster SC, Tsaliki E, et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature*. 2019;574 (7776):117–121. PMID: 31534227. doi:10.1038/s41586-019-1560-1
26. Neu J, Rushing J. Cesarean versus vaginal delivery: long-term infant outcomes and the hygiene hypothesis. *Clin Perinatol*. 2011;38(2):321–331. PMID: 21645799. doi:10.1016/j.clp.2011.03.008
27. Li Z, Yuan T. Neutrophil extracellular traps in adult diseases and neonatal bacterial infectious diseases: a review. *Heliyon*. 2024;10(1):e23559. PMID: 38173520. doi:10.1016/j.heliyon.2023.e23559
28. Zhou E, Wang K, Gu Y. Research Progress in Neonatal Alloimmune Thrombocytopenia: a Narrative Review. *Altern Ther Health Med*. 2023;29 (6):77–81. PMID: 37318890.
29. Dermyshi E, Wang Y, Yan C, et al. The "Golden Age" of probiotics: a systematic review and meta-analysis of randomized and observational studies in preterm infants. *Neonatology*. 2017;112(1):9–23. PMID: 28196365. doi:10.1159/000454668
30. O'Grady NP, Alexander M, Dellinger EP, et al. Guidelines for the prevention of intravascular catheter-related infections. The Hospital Infection Control Practices Advisory Committee, Center for Disease Control and Prevention, U.S. *Pediatrics*. 2002;110(5):e51. PMID: 12415057. doi:10.1542/peds.110.5.e51
31. Sharma D, Shastri S, Sharma P. Intrauterine growth restriction: antenatal and postnatal aspects. *Clin Med Insights Pediatr*. 2016;10:67–83. PMID: 27441006. doi:10.4137/CMPed.S40070
32. Steer P. The epidemiology of preterm labour. *BJOG Int J Obstet Gynaecol*. 2005;112(Suppl 1):1–3. PMID: 15715585. doi:10.1111/j.1471-0528.2005.00575.x
33. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. *Lancet*. 2008;371(9606):75–84. PMID: 18177778. doi:10.1016/S0140-6736(08)60074-4
34. Redman CW, Sargent IL. Immunology of pre-eclampsia. *Am J Reprod Immunol*. 2010;63(6):534–543. PMID: 20331588. doi:10.1111/j.1600-0897.2010.00831.x
35. Gestational Hypertension and Preeclampsia: ACOG Practice Bulletin Summary, Number 222. *Obstet Gynecol*. 2020;135(6):1492–1495. PMID: 32443077. doi:10.1097/AOG.0000000000003892
36. Kramer MS, Kakuma R. Optimal duration of exclusive breastfeeding. *Cochrane Database Syst Rev*. 2012;8:CD003517. PMID: 22895934. doi:10.1002/14651858.CD003517.pub2