TYPE Brief Research Report
PUBLISHED 04 April 2024
DOI 10.3389/frhem.2024.1341225

Constructing a novel clinical indicator model to predict the occurrence of thalassemia in pregnancy through machine learning algorithm

Yaoshui Long[†] and Wenxue Bai*[†]

Department of Clinical Laboratory, The Second People's Hospital of Jiangjin District, Chongqing, China

Check for updates

OPEN ACCESS

EDITED BY
Mariasanta Napolitano,
University of Palermo, Italy

REVIEWED BY
Richard Oscar Francis,
Columbia University, United States
Mohamed A. Yassin,
Qatar University, Qatar
Hakan Ayyildiz,
Istanbul University, Türkiye

*CORRESPONDENCE
Wenxue Bai
✉ 6680806@qq.com

[†]These authors have contributed equally to this work and share first authorship

RECEIVED 28 November 2023
ACCEPTED 14 March 2024
PUBLISHED 04 April 2024

CITATION
Long Y and Bai W (2024) Constructing a novel clinical indicator model to predict the occurrence of thalassemia in pregnancy through machine learning algorithm.
Front. Hematol. 3:1341225.
doi: 10.3389/frhem.2024.1341225

Thalassemia is one of the inherited hemoglobin disorders worldwide, resulting in ineffective erythropoiesis, chronic hemolytic anemia, compensatory hemopoietic expansion, hypercoagulability, etc., and when a mother carries the thalassemia gene, the child is more likely to have severe thalassemia. Furthermore, the economic and time costs of genetic testing for thalassemia prevent many thalassemia patients from being diagnosed in time. To solve this problem, we performed least absolute shrinkage and selection operator (LASSO) regression to analyze the correlation between thalassemia and blood routine indicators containing mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and red blood cell (RBC). We then built a nomogram to predict the occurrence of thalassemia, and receiver operating characteristic (ROC) curve was used to verify the prediction efficiency of this model. In total, we obtained 7,621 cases, including 847 thalassemia patients and 6,774 non-thalassemia. Among the 847 thalassemia patients, with a positivity rate of 67.2%, 569 cases were positive for α-thalassemia, and with a rate of 31.5%, 267 cases were positive for β-thalassemia. The remaining 11 cases were positive for both α- and β-thalassemia. Based on machine learning algorithm, we screened four optimal indicators, namely, MCV, MCH, RBC, and MCHC. The AUC value of MCV, MCH, RBC, and MCHC were 0.907, 0.906, 0.796, and 0.795, respectively. Moreover, the AUC value of the prediction model was 0.911. In summary, a novel and effective machine learning model was built to predict thalassemia, which functioned accurately, and may provide new insights for the early screening of thalassemia in the future.

KEYWORDS

thalassemia, machine learning, blood routine indicators, pregnancy, prediction

1 Introduction

Thalassemia, a prevalent monogenic disease worldwide, leads to hemolytic anemia due to impaired globin synthesis (1, 2). It is particularly prevalent in regions such as South Africa, the Middle East, and Southeast Asia, as well as in low- and middle-income areas like coastal cities in southern China and rural areas in western China. China bears the highest burden of thalassemia globally, with approximately 30 million individuals affected by thalassemia-related mutations and 3 million suffering from moderate to severe forms, posing significant challenges to families and society (3–5). Due to the autosomal recessive inheritance pattern of thalassemia, parents who are asymptomatic can still have children affected by thalassemia. When both parents carry the thalassemia gene, there is a substantial likelihood of their child developing severe thalassemia, which typically necessitates lifelong blood transfusions and is accompanied by various complications. This imposes a significant burden on affected children, their families, and society. Consequently, early detection of thalassemia during pregnancy assumes paramount importance. Existing approaches for thalassemia screening and diagnosis encompass osmotic fragility tests, assessment of red blood cell (RBC) smears, identification of inclusion bodies, evaluation of red blood cell indices, hemoglobin electrophoresis, high-performance liquid chromatography (HPLC), and genetic testing (6, 7). However, the cost and time associated with genetic testing often hinder timely diagnosis for many thalassemia patients. Conversely, blood routine indicators play a crucial role in the early identification of thalassemia due to the widespread availability of blood routine tests and the ability to distinguish different types of anemia based on red blood cell morphology (8, 9).

In this study, we analyzed the genetic test and blood routine results of 7,621 pregnant women being tested for thalassemia in the Jiangjin area of Chongqing from 2018 to 2022. Additionally, we employed a machine learning model to investigate the predictive value of blood routine indicators for thalassemia. The goal was to offer novel strategies for the early diagnosis, genetic counseling, and treatment of thalassemia in pregnant women within the Chongqing region.

2 Materials and methods

2.1 Patients

We conducted a retrospective study on thalassemia patients who were pregnant in the Jiangjin District, focusing on prenatal screening. From January 2018 to December 2022, blood routine and genetic tests for thalassemia were performed for the first time. After removing duplicates and cases with missing key information, a total of 7,621 cases were included in our analysis. This study received approval from the Medical Ethics Committee of The Second People's Hospital of Jiangjin District, Chongqing.

2.2 Blood routine index test

The Sysmex XN-1000 automated blood cell analysis system along with its accompanying reagents were utilized to measure various hematological parameters, including red blood cell count (RBC), hemoglobin level (Hb), hematocrit level (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and the standard deviation of red cell volume distribution width (RDW-SD).

2.3 Genetic testing for thalassemia

In our study, we employed polymerase chain reaction (PCR) in combination with diversion hybridization to detect various types of mutations and deletions associated with α-thalassemia and β-thalassemia. Specifically, we targeted three deletion types of α-thalassemia (i.e., –SEA, -α3.7 and -α4.2), three mutation types of α-thalassemia (CS, QS, and WS), as well as 17 mutation types of β-thalassemia [-28(A-G), -29(A-G), -30(T-C), -32(C-A), CD14/15 (+G), CD17(A-T), CD27/28(+C), CD31(-C), CD41/42(-TCTT), CD43(G-T), CD71/72(+A), IVS-I-1(G-T,G-A), IVS-I-5(G-C), IVS-II-654(C-T), βE(G-A), CAP(A-C, A-AAAC), and Int(T-G)].

2.4 Model establishment

Data collation was conducted using Microsoft Excel (RRID : SCR_016137). The R Project for Statistical Computing (RRID : SCR_001905) was utilized for model establishment, training, verification of factors associated with hematological indicators, and thalassemia prediction. The samples were randomly divided into two parts: a training set and a validation set. The training set consisted of Type 1 [IVS-II-654 (C-T), 55 cases], Type 2 (–SEA/αα, 186 cases), Type 3 (-α3.7/αα, 287 cases), Type 4 [CD17 (A-T), 95 cases], and Type 5 [CD41-42 (-TCTT), 62 cases]. To identify the optimal indicators for the prediction model, least absolute shrinkage and selection operator (LASSO) regression was performed via the R package "glmnet". The veen plot was visualized by the R package "ggVennDiagram". The predictive accuracy of the model was verified using the receiver operating characteristic (ROC) curve by the R package "pROC". Furthermore, a nomogram was constructed to establish the scoring criteria for the corresponding variables based on the coefficients of the LASSO regression model using the R package "rms". The selected variables were plotted on the variable axis, and a straight line was drawn to determine the score for each variable value. Using the training set, the blood routine data were imported, and the scores corresponding to each variable were assigned. The total score was calculated by summing the scores of all variables. These data were then inputted into the linear predictor to predict the risk of thalassemia.

3 Results

3.1 Positive genetic types of thalassemia

Among the 7,621 cases analyzed, a total of 847 were identified as positive for thalassemia, resulting in a positivity rate of 11.11%.

Specifically, 569 cases were positive for α-thalassemia, with a rate of 7.47%, while 267 cases were positive for β-thalassemia, with a rate of 3.50%. Furthermore, there were 11 cases that tested positive for both α- and β-thalassemia, as well as other combined genotypes, accounting for 0.14% of the cases. Compared to normal pregnant women, individuals with thalassemia exhibited decreased levels of hemoglobin (HGB), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and increased red blood cell distribution width-standard deviation (RDW-SD) and red blood cell distribution width-coefficient of variation (RDW-CV). These findings are consistent with the clinical manifestations of small cell hypochromic anemia. Notably, the changes observed in individuals with β-thalassemia and α- combined β-thalassemia genotypes were more pronounced than those seen in individuals with α-thalassemia (as shown in Table 1).

3.2 Analysis of the relevant coefficient of thalassemia

To identify the variables with the highest correlation to thalassemia, a LASSO prediction model was established using the data from the training set, which consisted of five thalassemia genotypes, namely, Type 1 [IVS-II-654 (C-T), 55 cases] (Figure 1A), Type 2 (–SEA/αα, 186 cases) (Figure 1B), Type 3 (-α3.7/αα, 287 cases) (Figure 1C), Type 4 [CD17 (A-T), 95 cases] (Figure 1D), and Type 5 [CD41-42 (-TCTT), 62 cases] (Figure 1E). Binomial deviation and Venn diagram analyses were employed to select the optimal variables (Figure 1F). As a result, four variables, namely, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and red blood cell count (RBC), were found to have the strongest correlation with thalassemia. To validate the predictive performance of these variables, receiver operating characteristic (ROC) curve analysis was conducted. The area under the curve (AUC) (Figure 1G) was used to interpret the results, with values greater than 0.5 and closer to 1 indicating higher effectiveness. It was observed that MCV and MCH exhibited a higher correlation with thalassemia compared to RBC and MCHC.

3.3 Establishment of a prediction model for thalassemia

Utilizing the correlation factors of MCV, MCH, RBC, and MCHC, a nomogram model (Figure 2A) was constructed, incorporating the scoring criteria for each variable derived from the coefficients of the LASSO regression model. The training set was selected to establish the scoring criteria for different variables. The receiver operating characteristic (ROC) curve analysis was performed to evaluate the predictive performance of MCV, MCH, RBC, and MCHC. The respective area under the curve (AUC) values were found to be 0.910, 0.909, 0.807, and 0.801

TABLE 1 The collected clinical information of all the data.

	AGE	RBC	Hb	HCT	MCV	MCH	MCHC	RDW-SD	RDW-CV	Gestation_period		
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	early_pregnancy	mid_pregnancy	late_pregnancy
-$\alpha^{3.7}/\alpha\alpha$ (N=287)	27.6 (5.11)	4.35 (0.447)	119 (12.0)	36.4 (3.39)	84.0 (4.89)	27.5 (2.11)	327 (11.1)	41.3 (5.09)	13.7 (2.16)	229 (79.8%)	47 (16.4%)	11 (3.8%)
$\_^{SEA}/\alpha\alpha$ (N=186)	27.1 (5.12)	5.03 (0.539)	109 (11.4)	34.6 (3.44)	69.0 (4.97)	21.8 (1.91)	316 (12.0)	37.3 (2.70)	15.5 (1.44)	146 (78.5%)	30 (16.1%)	10 (5.4%)
$\beta^{CD17(A\text{-}T)}/\beta^{N}$ (N=95)	27.8 (5.46)	4.91 (0.579)	98.9 (9.83)	31.3 (3.02)	64.0 (3.95)	20.2 (1.55)	316 (12.4)	36.0 (3.05)	16.7 (1.48)	74 (77.9%)	17 (17.9%)	4 (4.2%)
$\beta^{CD41\text{-}42(\text{-}TCTT)}/\beta^{N}$ (N=62)	26.7 (5.03)	4.77 (0.641)	99.8 (15.1)	31.4 (4.09)	66.1 (6.75)	21.1 (2.76)	318 (16.3)	37.8 (5.84)	16.8 (1.85)	50 (80.6%)	8 (12.9%)	4 (6.5%)
$\beta^{IVS\text{-}II\text{-}654(C\text{-}T)}/\beta^{N}$ (N=55)	26.8 (5.23)	4.80 (0.538)	100 (10.9)	31.2 (3.27)	65.4 (5.92)	21.0 (2.04)	320 (13.1)	36.4 (3.37)	16.5 (1.67)	45 (81.8%)	9 (16.4%)	1 (1.8%)
-$\alpha^{4.2}/\alpha\alpha$ (N=38)	26.0 (4.58)	4.34 (0.389)	117 (10.9)	36.0 (2.85)	82.9 (3.73)	27.0 (1.68)	326 (12.6)	40.1 (3.14)	13.4 (0.960)	28 (73.7%)	7 (18.4%)	3 (7.9%)

(Continued)

TABLE 1 Continued

	AGE	RBC	Hb	HCT	MCV	MCH	MCHC	RDW-SD	RDW-CV	Gestation_period		
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	early_pregnancy	mid_pregnancy	late_pregnancy
$\alpha^{CS}\alpha/\alpha\alpha$ (N=25)	26.3 (5.19)	4.34 (0.494)	116 (14.8)	35.5 (3.82)	82.0 (4.41)	26.7 (1.80)	325 (11.0)	40.0 (4.19)	13.7 (2.14)	18 (72.0%)	6 (24.0%)	1 (4.0%)
$\alpha^{WS}\alpha/\alpha\alpha$ (N=19)	26.3 (6.04)	4.35 (0.443)	123 (13.0)	37.1 (3.39)	85.5 (4.28)	28.3 (2.07)	332 (11.7)	41.0 (4.22)	13.2 (1.55)	18 (94.7%)	1 (5.3%)	0 (0%)
$\beta^{-28(A\text{-}G)}/\beta^{N}$ (N=16)	26.1 (3.45)	4.88 (0.447)	113 (8.22)	35.1 (2.84)	72.1 (2.37)	23.2 (1.08)	321 (11.4)	39.4 (3.38)	15.6 (1.35)	10 (62.5%)	4 (25.0%)	2 (12.5%)
$\beta^{CD43(G\text{-}T)}/\beta^{N}$ (N=11)	27.2 (4.62)	4.61 (0.514)	95.6 (8.92)	30.6 (2.53)	66.7 (5.85)	20.8 (1.97)	312 (6.59)	36.2 (3.88)	16.0 (2.13)	5 (45.5%)	3 (27.3%)	3 (27.3%)
$\beta^{E(G\text{-}A)}/\beta^{N}$ (N=9)	27.0 (5.61)	4.36 (0.323)	114 (7.83)	34.6 (2.14)	79.4 (2.88)	26.1 (1.05)	328 (5.03)	38.9 (2.31)	13.6 (0.743)	7 (77.8%)	2 (22.2%)	0 (0%)
$\alpha^{QS}\alpha/\alpha\alpha$ (N=8)	28.5 (3.16)	4.88 (0.602)	119 (13.2)	36.9 (4.75)	75.5 (1.71)	24.4 (1.33)	323 (13.3)	39.7 (3.33)	14.8 (1.50)	7 (87.5%)	0 (0%)	1 (12.5%)
β^{Cap}/β^{N} (N=7)	27.6 (5.80)	4.31 (0.419)	127 (13.7)	38.1 (3.41)	88.5 (3.64)	29.6 (1.91)	335 (13.0)	40.5 (2.40)	12.6 (0.660)	6 (85.7%)	1 (14.3%)	0 (0%)
$\beta^{CD71\text{-}72(+A)}/\beta^{N}$ (N=5)	23.8 (5.36)	4.46 (0.560)	95.0 (12.1)	30.2 (3.56)	67.8 (1.74)	21.4 (0.924)	315 (7.66)	41.9 (2.68)	18.6 (1.49)	1 (20.0%)	3 (60.0%)	1 (20.0%)
$\beta^{-29(A\text{-}G)}/\beta^{N}$ (N=3)	23.7 (5.86)	4.59 (0.194)	103 (1.53)	32.8 (0.404)	71.6 (2.46)	22.5 (1.34)	315 (8.08)	36.0 (1.30)	14.2 (0.346)	3 (100%)	0 (0%)	0 (0%)
$\beta^{27\text{-}28(A\text{-}G)}/\beta^{N}$ (N=3)	27.3 (6.43)	4.64 (0.977)	99.7 (7.57)	31.4 (2.98)	69.0 (9.43)	22.0 (3.44)	318 (7.21)	42.1 (12.8)	17.5 (2.24)	2 (66.7%)	1 (33.3%)	0 (0%)
$\_^{SEA}/\alpha\alpha/\beta^{CD41\text{-}42}/\beta^{N}$ (N=3)	28.3 (8.02)	4.58 (0.448)	103 (3.51)	32.6 (0.462)	71.5 (7.02)	22.7 (2.60)	317 (6.66)	45.3 (9.83)	18.5 (4.90)	3 (100%)	0 (0%)	0 (0%)
$\_^{SEA}/\text{-}\alpha^{3.7}$ (N=2)	21.5 (3.54)	4.63 (0.523)	82.5 (9.19)	26.7 (0.891)	57.8 (4.54)	17.8 (0.0141)	309 (24.0)	43.3 (1.20)	25.3 (1.86)	2 (100%)	0 (0%)	0 (0%)
$\_^{SEA}/\text{-}\alpha^{3.7}/\beta^{CD17}/\beta^{N}$ (N=1)	34.0 (NA)	6.07 (NA)	100 (NA)	32.2 (NA)	53.0 (NA)	16.5 (NA)	311 (NA)	31.7 (NA)	20.2 (NA)	1 (100%)	0 (0%)	0 (0%)
$\_^{SEA}/\alpha\alpha/\beta^{CD17}/\beta^{N}$ (N=1)	28.0 (NA)	4.94 (NA)	114 (NA)	35.1 (NA)	71.1 (NA)	23.1 (NA)	325 (NA)	33.1 (NA)	13.1 (NA)	1 (100%)	0 (0%)	0 (0%)
$\text{-}\alpha^{3.7}/\text{-}\alpha^{4.2}$ (N=1)	18.0 (NA)	4.61 (NA)	97.0 (NA)	32.1 (NA)	69.6 (NA)	21.0 (NA)	302 (NA)	35.1 (NA)	14.6 (NA)	0 (0%)	0 (0%)	1 (100%)

(Continued)

TABLE 1 Continued

	AGE	RBC	Hb	HCT	MCV	MCH	MCHC	RDW-SD	RDW-CV	Gestation_period		
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	early_pregnancy	mid_pregnancy	late_pregnancy
$-\alpha^{3.7}/\alpha\alpha/\beta^{-28(A\text{-}G)}/\beta^{N}$ (N=1)	19.0 (NA)	4.07 (NA)	104 (NA)	31.9 (NA)	78.4 (NA)	25.6 (NA)	326 (NA)	41.5 (NA)	14.6 (NA)	0 (0%)	1 (100%)	0 (0%)
$-\alpha^{3.7}/\alpha\alpha/\beta^{CD41\text{-}42}/\beta^{N}$ (N=1)	21.0 (NA)	4.91 (NA)	105 (NA)	32.1 (NA)	65.4 (NA)	21.4 (NA)	327 (NA)	34.6 (NA)	15.0 (NA)	1 (100%)	0 (0%)	0 (0%)
$-\alpha^{3.7}/\alpha^{CS}\alpha$ (N=1)	33.0 (NA)	5.54 (NA)	125 (NA)	40.8 (NA)	73.6 (NA)	22.6 (NA)	306 (NA)	42.1 (NA)	16.3 (NA)	0 (0%)	1 (100%)	0 (0%)
$-\alpha^{3.7}//\alpha\alpha/\beta^{CD17}/\beta^{N}$ (N=1)	22.0 (NA)	5.00 (NA)	101 (NA)	32.1 (NA)	64.2 (NA)	20.2 (NA)	315 (NA)	34.8 (NA)	15.7 (NA)	1 (100%)	0 (0%)	0 (0%)
$-\alpha^{4.2}/\alpha\alpha/\beta^{CD17}/\beta^{N}$ (N=1)	26.0 (NA)	5.07 (NA)	107 (NA)	32.6 (NA)	64.3 (NA)	21.1 (NA)	328 (NA)	35.7 (NA)	16.1 (NA)	1 (100%)	0 (0%)	0 (0%)
$-\alpha^{4.2}/\alpha^{CS}\alpha$ (N=1)	23.0 (NA)	4.15 (NA)	104 (NA)	33.0 (NA)	79.5 (NA)	25.1 (NA)	315 (NA)	40.4 (NA)	14.1 (NA)	1 (100%)	0 (0%)	0 (0%)
$-\alpha^{4.2}/\alpha^{WS}\alpha$ (N=1)	23.0 (NA)	5.23 (NA)	128 (NA)	39.6 (NA)	75.7 (NA)	24.5 (NA)	323 (NA)	37.9 (NA)	14.1 (NA)	1 (100%)	0 (0%)	0 (0%)
$\alpha^{WS}\alpha/\alpha\alpha/\beta^{CD41\text{-}42}/\beta^{N}$(N=1)	19.0 (NA)	5.07 (NA)	97.0 (NA)	31.4 (NA)	61.9 (NA)	19.1 (NA)	309 (NA)	31.1 (NA)	14.7 (NA)	1 (100%)	0 (0%)	0 (0%)
$\beta^{CD14\text{-}15(+G)}/\beta^{N}$ (N=1)	20.0 (NA)	3.64 (NA)	93.0 (NA)	30.5 (NA)	83.8 (NA)	25.5 (NA)	305 (NA)	43.1 (NA)	14.4 (NA)	0 (0%)	0 (0%)	1 (100%)
Fusion gene/αα (N=1)	30.0 (NA)	3.76 (NA)	108 (NA)	33.6 (NA)	89.4 (NA)	28.7 (NA)	321 (NA)	45.3 (NA)	14.0 (NA)	0 (0%)	1 (100%)	0 (0%)
Normal (N=6,774)	27.0 (4.82)	4.07 (0.457)	121 (13.1)	36.2 (3.55)	89.6 (5.50)	30.0 (2.36)	335 (11.3)	42.3 (3.84)	13.0 (1.47)	7,367 (83.7%)	1,275 (14.5%)	163 (1.9%)
Overall (N=7,621)	27.0 (4.85)	4.12 (0.498)	120 (13.5)	36.1 (3.62)	88.3 (7.31)	29.5 (3.00)	334 (12.1)	42.0 (4.06)	13.2 (1.65)	8,029 (83.2%)	1,417 (14.7%)	206 (2.1%)

"NA" indicates that SD cannot be counted because there is only one data.

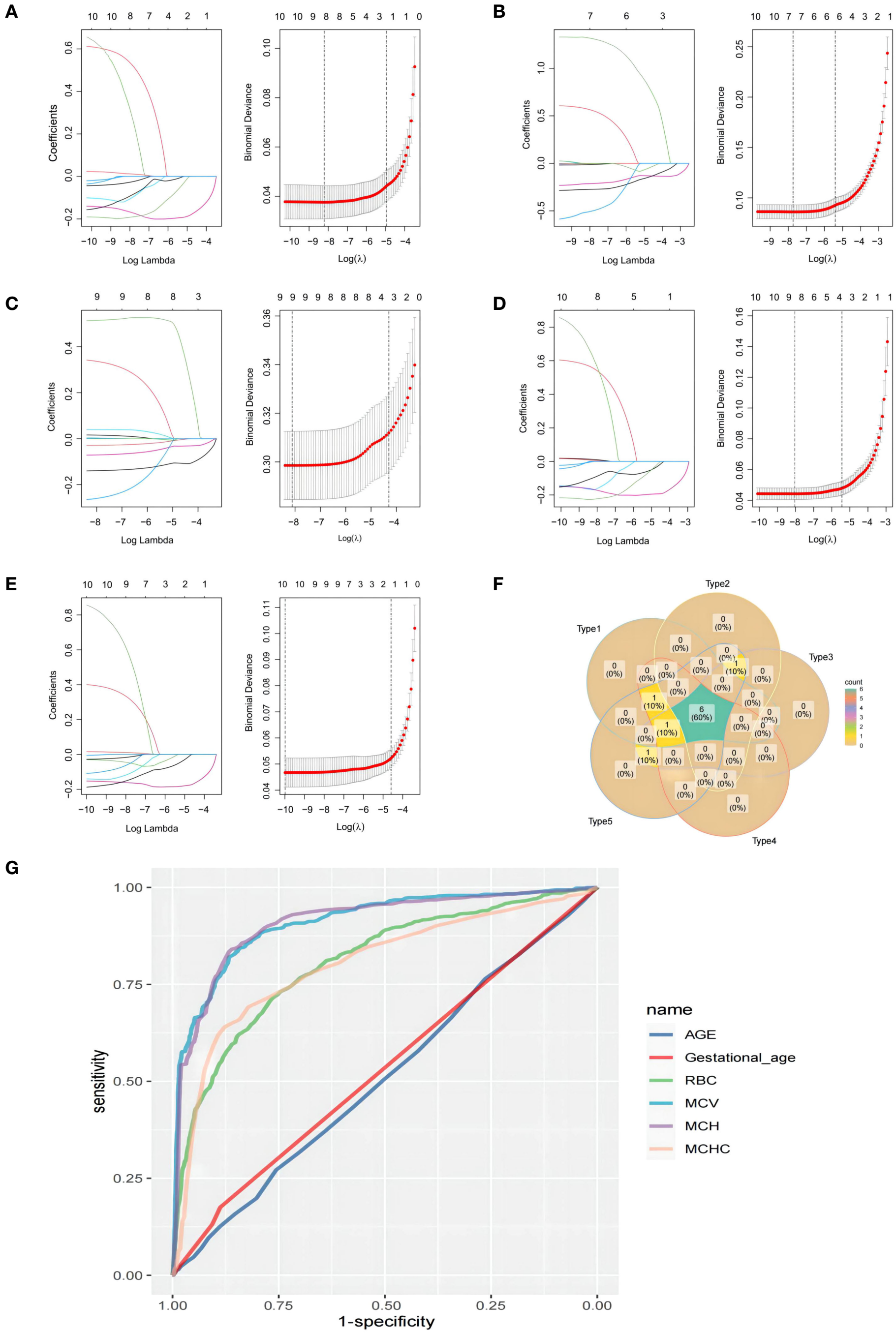

FIGURE 1
Least absolute shrinkage and selection operator (LASSO) prediction model was established using the data from the training set, which consisted of five thalassemia genotypes: **(A)** Type 1 [IVS-II-654(C-T), 55 cases], **(B)** Type 2 (–SEA/αα, 186 cases), **(C)** Type 3 (-$\alpha^{3.7}$/αα, 287 cases), **(D)** Type 4 [CD17 (A-T), 95 cases], and **(E)** Type 5 [CD41-42(-TCTT), 62 cases]. **(F)** The veen plot of the LASSO result intersections among the five types. **(G)** The receiver operating characteristic (ROC) curve for the predictive value of each factor.

(Figures 2B–E). Furthermore, the AUC value of the overall model was determined to be 0.913 (Figure 2F). These results demonstrate that the prediction model effectively enhances the predictive power of these variables and reduces errors associated with using a single index alone.

3.4 Optimization of thalassemia prediction model

After observing the satisfactory performance of the model in the training set, we proceeded to fine-tune the variables to accommodate

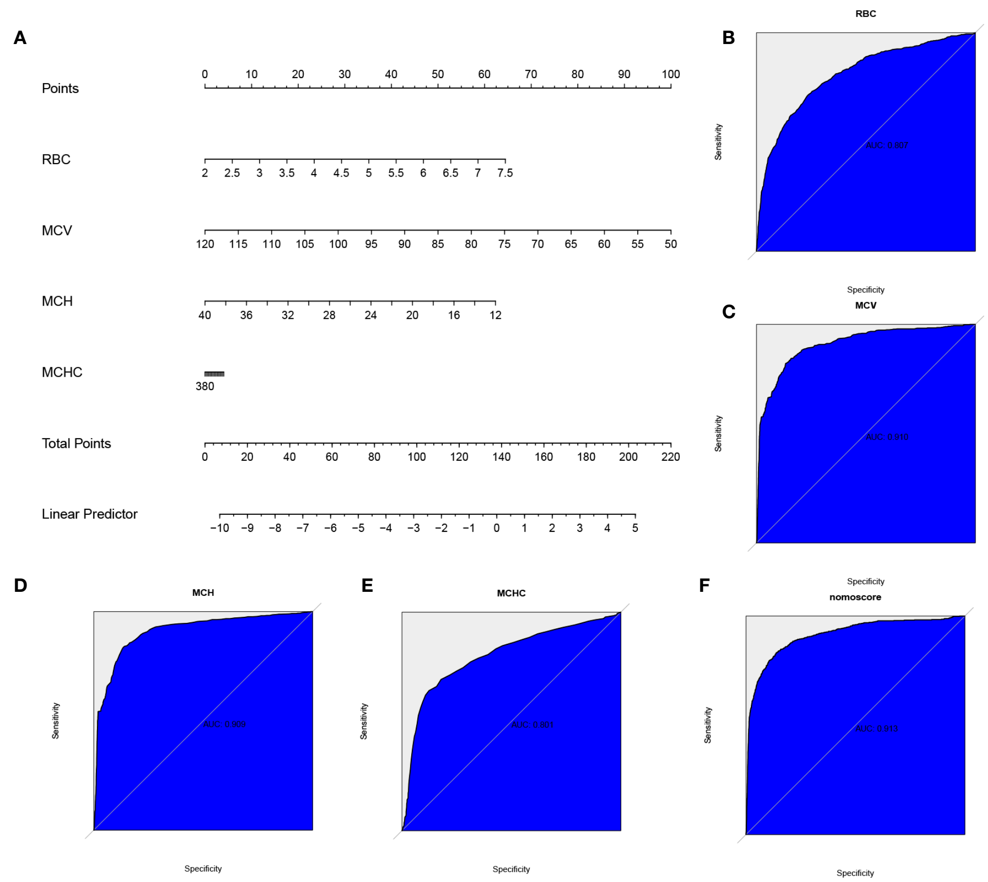

FIGURE 2
The nomogram and predictive efficiency of this model in the training set. **(A)** Construction of a predictive nomogram. The receiver operating characteristic (ROC) curve standing for the predictive efficiency of the correlation factors: **(B)** red blood cell (RBC), **(C)** mean corpuscular volume (MCV), **(D)** mean corpuscular hemoglobin (MCH), **(E)** mean corpuscular hemoglobin concentration (MCHC), and **(F)** the model.

more complex data in the testing set (Figure 3A). As a result, the AUC values for MCV, MCH, RBC, and MCHC were found to be 0.907, 0.906, 0.796, and 0.795, respectively (Figures 3B–E). Notably, the overall model exhibited an AUC value of 0.911, indicating a strong predictive effect (Figure 3F). In the validation group, the addition of normal data for thalassemia and other genotypes led to a reduction in AUC. This adjustment was made to bring the model closer to real-world scenarios and enhance its reliability.

4 Discussion

The rapid advancement of artificial intelligence (AI) technology has garnered significant attention worldwide, particularly in the realm of disease diagnosis. Machine learning and other related technologies have been widely explored to aid in this process. In comparison to manual diagnosis, computer-based diagnostic methods offer increased accuracy and efficiency, effectively reducing the misdiagnosis rate. Consequently, these methods enable more effective disease diagnoses at a lower cost (10–12). By harnessing the power of big data and clinical information, machine learning techniques have greatly improved the accuracy and efficiency of clinical diagnoses. This progress has propelled laboratory medicine toward precision medicine and intelligent testing (13, 14). Applying machine learning to the early diagnosis of thalassemia, for instance, allows for the prediction of thalassemia types with higher severity using a blood sample with high prevalence. This approach can effectively reduce the cost and time required for genetic identification of thalassemia, enabling healthcare providers to initiate early treatment measures and offer prompt genetic counseling to pregnant women with thalassemia. Such efforts are important in reducing the birth rate of children with moderate and severe thalassemia and conducting population health surveys in middle and low-income areas.

In this study, a total of 7,621 pregnant women underwent blood routine and thalassemia detection. Among them, 569 cases of α-thalassemia were identified, yielding a positive rate of 7.47%. Additionally, 267 cases of β-thalassemia were detected, with a positive rate of 3.5%. Furthermore, 11 cases of α- combined with β-thalassemia were observed, resulting in a positive rate of 0.14%. These prevalence rates were found to be lower compared to coastal cities such as Guangdong, particularly in relation to α thalassemia, indicating a significant difference as compared to previous studies (4, 15). To mitigate the cost and time associated with genetic diagnosis of thalassemia, a machine learning model was constructed in this study. The model employed routine blood test

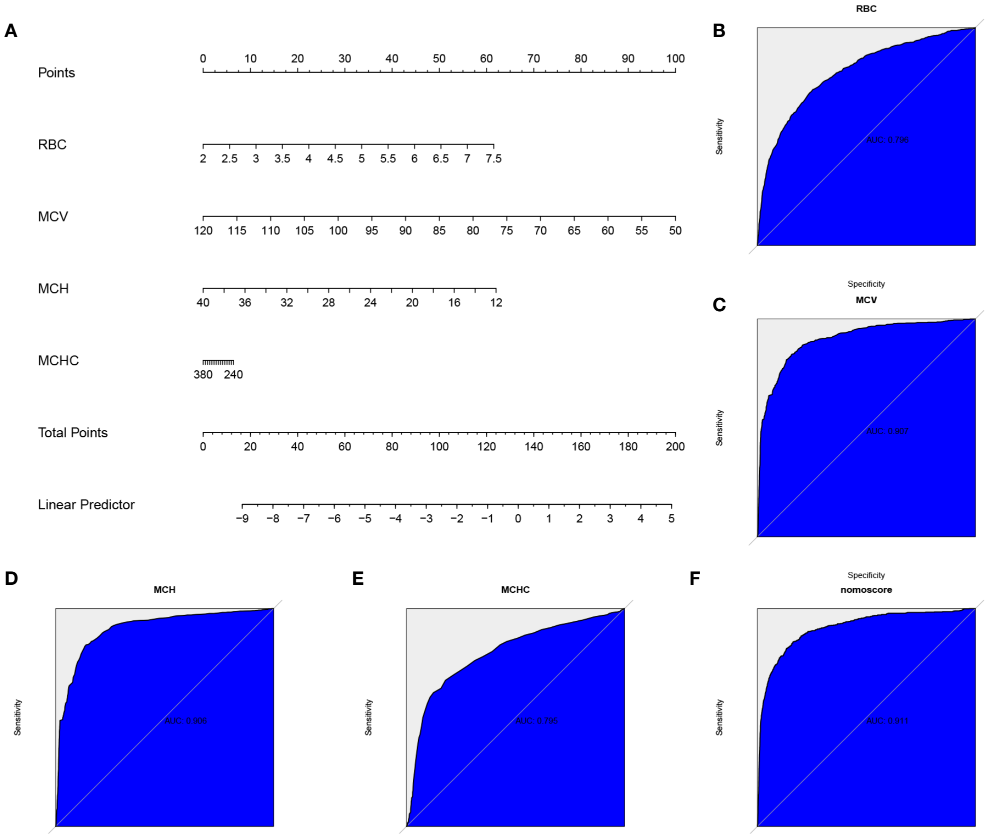

FIGURE 3
The nomogram and predictive efficiency of this model in the testing set. **(A)** Construction of a predictive nomogram. The ROC curve standing for the predictive efficiency of the correlation factors: **(B)** red blood cell (RBC), **(C)** mean corpuscular volume (MCV), **(D)** mean corpuscular hemoglobin (MCH), **(E)** mean corpuscular hemoglobin concentration (MCHC), and **(F)** the model.

results for thalassemia prediction. Among the five genotypes with the highest incidence of thalassemia [α3.7/αα, –SEA/αα, CD17 (A-T), CD41-42 (-TCTT), and IVS-II-654 (C-T)], the four coefficients that exhibited the highest correlation with thalassemia were selected: MCV, MCH, RBC, and MCHC. Notably, the prediction performance of MCV and MCH demonstrated superior results. These findings align with previous studies, which have consistently reported reduced MCV and MCH levels across almost all types of thalassemia (15–17).

The prediction model was developed using four variables: MCV, MCH, RBC, and MCHC. The model's predictive performance was assessed using ROC analysis, yielding AUC values of 0.910, 0.909, 0.807, and 0.801, respectively. These results indicate significant diagnostic value for thalassemia detection. The overall AUC value of the model was found to be 0.913, surpassing the individual variables in terms of prediction accuracy. This result demonstrates the successful construction of the model. To further optimize the model for complex clinical data, adjustments were made to each variable. Subsequently, the entire dataset was utilized for verification purposes. The obtained AUC values for MCV, MCH, RBC, and MCHC were 0.907, 0.906, 0.796, and 0.795, respectively. These values were slightly lower than those obtained during the training phase. However, the model still exhibited a good prediction effect, with an overall AUC value of 0.911. This minor decrease in AUC can be attributed to the inclusion of data from normal pregnant women. Their physiological anemia can interfere with the model, but incorporating such data ensures that the predictive performance of the model aligns more closely with clinical reality, thereby holding significant clinical significance.

It is worth mentioning that the main positive data in the model primarily consist of common thalassemia genotypes, such as -α3.7/αα, –SEA/αα, and CD17 (A-T). Consequently, the model's predictions may be more accurate for these genotypes. Furthermore, it is crucial to note that unlike the findings of Khan et al. and El-Beshlawy et al. (18, 19), the current model's predictive capability is limited when it comes to β-thalassemia in regions such as Africa, the Middle East, and other areas, as it primarily focuses on Asia, where α-thalassemia is prevalent. Additionally, since the study subjects predominantly consist of pregnant women, the model may face challenges in accurately predicting thalassemia in men and children. Another limitation of the model is its current inability to differentiate between varying degrees of thalassemia severity, such as severe, moderate, or characteristic types. Furthermore, research by Ferih et al. (20) provides us with a good idea that our model may face challenges in distinguishing between thalassemia and iron-deficiency anemia, as the diagnosis of iron-deficiency anemia often requires additional hematological indicators that may not be fully incorporated into the current model, which need further study.

5 Summary

This study aimed to predict thalassemia in pregnancy by means of economical and rapid blood routine detection by establishing a data prediction model. Through the analysis of 7,621 cases, MCV, MCH, RBC, and MCHC were selected as high correlation indicators. Subsequently, a machine learning prediction model was constructed, incorporating these four indicators as variables, and the results were verified using ROC analysis. The AUC values for MCV, MCH, RBC, and MCHC were 0.907, 0.906, 0.796, and 0.795, respectively. In particular, the prediction model achieved an AUC value of 0.911, demonstrating its effectiveness in thalassemia prediction and provided a novel strategy for the early screening of thalassemia.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Ethics Committee of The Second People's Hospital of Jiangjin District, Chongqing. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from primarily isolated as part of your previous study for which ethical approval was obtained. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

YL: Writing – original draft, Writing – review & editing. WB: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Science-Health Joint Medical Scientific Research Project of Chongqing (2022MSXM092).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Kattamis A, Kwiatkowski JL, Aydinok Y. thalassemia. *Lancet.* (2022) 399:2310–24. doi: 10.1016/S0140-6736(22)00536-0

2. Taher AT, Weatherall DJ, Cappellini MD. thalassemia. *Lancet.* (2018) 391:155–67. doi: 10.1016/S0140-6736(17)31822-6

3. Mo D, Zheng Q, Xiao B, Li L. Predicting thalassemia using deep neural network based on red blood cell indices. *Clin Chim Acta.* (2023) 543:117329. doi: 10.1016/j.cca.2023.117329

4. Wang WD, Hu F, Zhou DH, Gale RP, Lai YR, Yao HX, et al. thalassemia in China. *Blood Rev.* (2023) 60:101074. doi: 10.1016/j.blre.2023.101074

5. Weatherall DJ. The evolving spectrum of the epidemiology of thalassemia. *Hematol Oncol Clin North Am.* (2018) 32:165–75. doi: 10.1016/j.hoc.2017.11.008

6. Viprakasit V, Ekwattanakit S. Clinical classification, screening and diagnosis for thalassemia. *Hematol Oncol Clin North Am.* (2018) 32:193–211. doi: 10.1016/j.hoc.2017.11.006

7. Achour A, Koopmann TT, Baas F, Harteveld CL. The evolving role of next-generation sequencing in screening and diagnosis of hemoglobinopathies. *Front Physiol.* (2021) 12:686689. doi: 10.3389/fphys.2021.686689

8. Xu M, Lin G, Dong Z, Wang Q, Ma L, Su J. Logistic-Nomogram model based on red blood cell parameters to differentiate thalassemia trait and iron deficiency anemia in southern region of Fujian Province, China. *J Clin Lab Anal.* (2023) 37:e24940. doi: 10.1002/jcla.24940

9. Zheng S, Li Q, Ou T, Li Y, Wu S. Clinical performance study of a new fully automated red blood cell permeability fragility analyzer. *J Healthc Eng.* (2022) 2022:5642907. doi: 10.1155/2022/5642907

10. Yuan Y, Shi C, Zhao H. Machine learning-enabled genome mining and bioactivity prediction of natural products. *ACS Synth Biol.* (2023) 12(9):2650–62. doi: 10.1021/acssynbio.3c00234

11. Othman NA, Azhar MAAS, Damanhuri NS, Mahadi IA, Abbas MH, Shamsuddin SA, et al. Optimization of identifying insulinaemic pharmacokinetic parameters using artificial neural network. *Comput Methods Programs Biomed.* (2023) 236:107566. doi: 10.1016/j.cmpb.2023.107566

12. Bar-On M, Baharav S, Katzir Z, Mirelman A, Sosnik R, Maidan I. Task-related reorganization of cognitive network in Parkinson's disease using electrophysiology. *Mov Disord.* (2023) 38(11):2031–40. doi: 10.1002/mds.29571

13. Yuan W, Zhi W, Ma L, Hu X, Wang Q, Zou Y, et al. Neural oscillation disorder in the hippocampal CA1 region of different Alzheimer's disease mice. *Curr Alzheimer Res.* (2023) 20(5):350–9. doi: 10.2174/1567205020666230808122643

14. Rabbani N, Kim GYE, Suarez CJ, Chen JH. Applications of machine learning in routine laboratory medicine: Current state and future directions. *Clin Biochem.* (2022) 103:1–7. doi: 10.1016/j.clinbiochem.2022.02.011

15. Xian J, Wang Y, He J, Li S, He W, Ma X, et al. Molecular epidemiology and hematologic characterization of thalassemia in Guangdong Province, Southern China. *Clin Appl Thromb Hemost.* (2022) 28:10760296221119807. doi: 10.1177/10760296221119807

16. Zhuang J, Jiang Y, Wang Y, Zheng Y, Zhuang Q, Wang J, et al. Molecular analysis of α-thalassemia and β-thalassemia in Quanzhou region Southeast China. *J Clin Pathol.* (2020) 73:278–82. doi: 10.1136/jclinpath-2019-206179

17. Chen P, Lin WX, Li SQ. THALASSEMIA in ASIA 2021: thalassemia in Guangxi Province, People's Republic of China. *Hemoglobin.* (2022) 46:33–5. doi: 10.1080/03630269.2021.2008960

18. Khan AM, Al-Sulaiti AM, Younes S, Yassin M, Zayed H. The spectrum of beta-thalassemia mutations in the 22 Arab countries: a systematic review. *Expert Rev Hematol*. (2021) 14:109–22. doi: 10.1080/17474086.2021.1860003

19. El-Beshlawy A, Dewedar H, Hindawi S, Alkindi S, Tantawy AA, Yassin MA, et al. Management of transfusion-dependent β-thalassemia (TDT): Expert insights and practical overview from the Middle East. *Blood Rev*. (2024) 63:101138. doi: 10.1016/j.blre.2023.101138

20. Ferih K, Elsayed B, Elshoeibi AM, Elsabagh AA, Elhadary M, Soliman A, et al. Applications of artificial intelligence in thalassemia: a comprehensive review. *Diagnost (Basel)*. (2023) 13:1551. doi: 10.3390/diagnostics13091551