Original Article

Development and validation of an explainable machine learning-based prediction model for primary Kawasaki disease complicated with coronary artery aneurysms

Zixia Song[1,2], Hongjun Ming[1], Bin Liu[2], Dong Liu[2]

[1]Department of Pediatrics, Beijing Anzhen Nanchong Hospital, Capital Medical University (Nanchong Central Hospital), Nanchong, China; [2]Department of Pediatrics, The Affiliated Hospital of Southwest Medical University, Sichuan Clinical Research Center for Birth Defects, Luzhou, China

Contributions: (I) Conception and design: Z Song, D Liu, B Liu; (II) Administrative support: B Liu, D Liu; (III) Provision of study materials or patients: Z Song, H Ming; (IV) Collection and assembly of data: Z Song; (V) Data analysis and interpretation: Z Song, D Liu, B Liu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Bin Liu, PhD; Dong Liu, PhD. Department of Pediatrics, The Affiliated Hospital of Southwest Medical University, Sichuan Clinical Research Center for Birth Defects, No. 8, Section 2, Kangcheng Road, Jiangyang District, Luzhou 646000, China. Email: 1040796435@qq.com; lblyfy@126.com.

Background: Kawasaki disease (KD) can lead to coronary artery aneurysms (CAA) in approximately 1 in 5 untreated children despite intravenous immunoglobulin (IVIG) therapy in the acute phase. The aim of this study is to develop and validate an explainable machine learning (ML)-based prediction model for CAA in KD.

Methods: This study retrospectively analyzed the clinical data of children diagnosed with primary KD at Nanchong Central Hospital, Sichuan Province between 2015 and 2023. Six models, including support vector machine (SVM), K-nearest neighbors (KNN), least absolute shrinkage and selection operator (Lasso), extreme gradient boosting (XGBoost), random forest (RF), and multilayer perceptron (MLP), based on ML algorithms were developed. The model with optimal performance was validated and the explainable SHapley Additive exPlanations (SHAP) analysis was used.

Results: A total of 327 children diagnosed with KD were included in the training set and validation set. Receiver operator characteristic curve analysis showed that XGBoost based model exhibited an optimal performance among the six models. Moreover, for a given CAA positive sample, the sum of the SHAP values of all variables of XGBoost represented the individual deviation from the mean predicted from the entire dataset.

Conclusions: The XGBoost algorithm-based explainable model might be used to predict the occurrence of CAA in children with KD.

Keywords: Kawasaki disease (KD); coronary artery aneurysm (CAA); machine learning (ML); prediction model

Submitted Sep 11, 2024. Accepted for publication Feb 10, 2025. Published online Feb 25, 2025.
doi: 10.21037/tp-24-359
View this article at: https://dx.doi.org/10.21037/tp-24-359

Introduction

Kawasaki disease (KD) is a systemic self-limiting vasculitis that can affect multiple organ systems. In addition to its typical clinical manifestations such as fever, rash, conjunctival injection, strawberry tongue, and cervical lymphadenopathy, some patients develop coronary artery lesions (CALs) and other cardiovascular complications. Among these, CAL is the most serious and is currently the leading cause of acquired heart disease in China (1,2). Once coronary artery aneurysms (CAA) occur, they can lead to cardiac events, including myocardial infarction and sudden cardiac death (3-5). CAA is a pathological manifestation

of localized coronary artery dilatation, which has a far greater impact on the body than simple CALs. CAA causes hemodynamic abnormalities, significantly increases the risk of thrombosis, and can lead to myocardial infarction and other major cardiovascular events, making recovery difficult. Currently, there are few predictive models for CAA. Therefore, establishing an explainable model for CAA is critical to prevention and disease management.

Currently, the prediction of CAA in children with KD is mainly based on conventional statistical models (6-8). However, conventional models often struggle with the high dimensionality and numerous feature characteristics of KD patient data, which can compromise both predictive accuracy and practical applicability. In addition, these models may not adequately account for the intricate interplay between multiple clinical and laboratory variables, resulting in less precise predictions. Existing models typically do not offer personalized predictions for individual patients, which is critical for tailoring optimal KD treatment strategies. These constraints collectively hinder the overall performance of current methods in terms of both predictive accuracy and practical utility (6-8). Several models have shown good performance in prediction for the occurrence of CAA (9,10). Studies in China have established prediction models for CALs using machine learning (ML) to improve the prediction performance of CAL (11,12). Although some models have shown good performance in prediction for CAA (9,10), they often do not fully utilize advanced ML techniques. These advanced techniques, such as deep learning and ensemble methods, can provide more robust and accurate predictions by better handling high-dimensional data and extracting complex patterns (13). Without rigorous validation, the reliability and applicability of models in real-world clinical settings remain uncertain. Some studies have included relatively small numbers of cases, particularly for positive samples (patients with CAA), which may affect the statistical power and reliability of the models. Studies that focus on a limited set of variables may not capture the full spectrum of factors contributing to the development of CAA. A more comprehensive approach that considers a wider range of variables could improve prediction accuracy.

Therefore, the current study aimed to develop and validate an explainable ML-based prediction model for CAA in children with KD using advanced techniques and a larger, more diverse set. We present this article in accordance with the TRIPOD reporting checklist (available at https://tp.amegroups.com/article/view/10.21037/tp-24-359/rc).

Highlight box

Key findings

- This study developed and validated an explainable model for prediction of coronary artery aneurysms (CAA) in Kawasaki disease (KD).

What is known and what is new?

- KD can lead to CAA in approximately 1 in 5 untreated children despite intravenous immunoglobulin therapy in the acute phase.
- The extreme gradient boosting (XGBoost) algorithm-based explainable model might be used to predict the occurrence of CAA in children with KD.

What is the implication, and what should change now?

- The XGBoost algorithm was used to develop an explainable prediction model for CAA in KD based on age, prealbumin, absolute basophil count, eosinophil percentage, C-reactive protein, and monocyte percentage. These factors provide interpretable prediction results that enable physicians to understand the model's decision-making process, increase their confidence in the model, and effectively use machine learning model predictions in clinical decision making to provide more accurate and personalized prevention and treatment plans for patients.

Methods

Study design and population

This study retrospectively analyzed the clinical data of children who were primarily diagnosed with KD, at Nanchong Central Hospital, Sichuan Province between January 1, 2015, and October 30, 2023. The KD collected in this study was diagnosed in accordance with the 2004 American Heart Association (AHA) guidelines for KD (14). After admission, 2 g/kg intravenous immunoglobulin (IVIG) was given. For patients with IVIG resistance, 2 g/kg IVIG was given again. All collected blood sample data were obtained within 24 hours prior to the diagnosis of KD. Once KD was diagnosed, the patients were promptly treated with a regimen of IVIG and aspirin. Color Doppler ultrasound was collected within 1 day before the first dose of IVIG. Inclusion criteria were: (I) primary diagnosed KD; (II) age ≤14 years old; (III) received IVIG treatment. The exclusion criteria were: (I) children with congenital heart disease, a history of rheumatic disease, or other cardiovascular diseases that may involve coronary artery damage; (II) recurrent cases of KD. The study was conducted in accordance with the Declaration of Helsinki (as

revised in 2013). This study was approved by the Medical Ethics Committee of Nanchong Central Hospital [No. 2024(024)]. This article is a retrospective study. Therefore, the Medical Ethics Committee of Nanchong Central Hospital waived the requirement to obtain distinct written informed consent from the patients.

Data collection and definitions

Basic information (raw data are available at: https://cdn.amegroups.cn/static/public/tp-24-359-1.xlsx), such as age (in months), sex, height and weight; clinical features, such as conjunctival hyperemia, rash, cervical lymphadenopathy, strawberry tongue, perianal changes, swelling or peeling of hands and feet, duration of fever before IVIG treatment, and whether it was complete KD were collected; laboratory indicators, such as C-reactive protein (CRP), white blood cell (WBC) count, neutrophil percentage (NEU), monocyte percentage (MONO), erythrocyte sedimentation rate (ESR), albumin (ALB), hematocrit (HCT), platelet count (PLT), alanine aminotransferase (ALT), aspartate aminotransferase (AST), hemoglobin (Hb), etc., as well as IVIG use and response and coronary artery echocardiographic results. Definitions of coronary artery aneurysm classification based on Z-score are shown in Table S1.

Statistical analysis

Data with more than 20% missing values were deleted. Remaining missing values were imputed using the random forest (RF) model. Data were processed and analyzed using R software (version 4.2.2) and Stata software (version 15.0). Categorical data were expressed as frequencies (percentages), and continuous data were expressed as medians (interquartile ranges). Chi-squared tests were used for categorical data analysis, *t*-tests for normally distributed continuous data with equal variances, and Mann-Whitney *U* tests for non-normally distributed or unequal variance data. Subjects were randomly divided into training set and validation set (7:3), and the Synthetic Minority Over-sampling Technique (SMOTE) was used to correct for class imbalance in the training set (data are available at https://cdn.amegroups.cn/static/public/tp-24-359-2.xlsx). Univariate logistic regression analysis was used to select variables. Six algorithms were used to select important variables: support vector machine (SVM), K-nearest neighbors (KNN), least absolute shrinkage and selection operator (Lasso), extreme gradient boosting (XGBoost), RF and multilayer perceptron (MLP), and tenfold cross-validation were performed. The optimal model was validated for discrimination, calibration, clinical net benefit, and classification performance using receiver operating characteristic (ROC) curves, calibration curves, decision curve analysis (DCA), and Precision-Recall (PR) curves. To define the contribution of each variable in individual CAA prediction, the SHapley Additive exPlanations (SHAP) Tree framework was applied for the optimal model. The SHAP value integrates both the effect per se of a given biological variable and the effects of this variable in interaction with other biological parameters (*Figure 1*). The P value <0.05 was considered statistically significant.

Results

Baseline characteristics

A total of 327 children were included in the study and randomly assigned to a training set (n=228) and a validation set (n=99) in a 7:3 ratio. In the training set, there were 26 cases of CAA (11.4%), 82 males (36.0%), and 49 cases of incomplete KD (IKD) (21.5%). In the validation set, there were 10 cases of CAA (10.1%), 40 males (40.4%), and 16 cases of IKD (16.2%). There were no statistically significant differences in baseline characteristics between the training and validation sets (all P>0.05) (*Table 1*). The SMOTE was used to handle a class imbalance in the training set, variables for ML are presented in *Table 2*.

Model development and validation

The 30 variables selected by univariate logistic regression analysis were further screened by using six ML algorithms: SVM, KNN, Lasso, XGBoost, RF, MLP. Each algorithm selected important variables for modeling, followed by tenfold cross-validation. Receiver operator characteristic curve analysis showed that XGBoost based model exhibits an optimal performance among the six models, XGBoost [training set: area under the curve (AUC) =0.982; validation set: AUC =0.958], SVM (1.000; 0.810), KNN (0.985; 0.734), Lasso (0.981; 0.729), RF (0.999; 0.892), and MLP (0.984; 0.719) (*Figure 2*). The XGBoost had the best performance. The PR curve plot also showed that XGBoost had a good classification performance in the validation set (*Figure 3*). The calibration curves plot of XGBoost revealed a good predictive accuracy between the actual probability and predicted probability in the validation set (P=0.63) (*Figure 4*). The clinical DCA showed that XGBoost had a good net

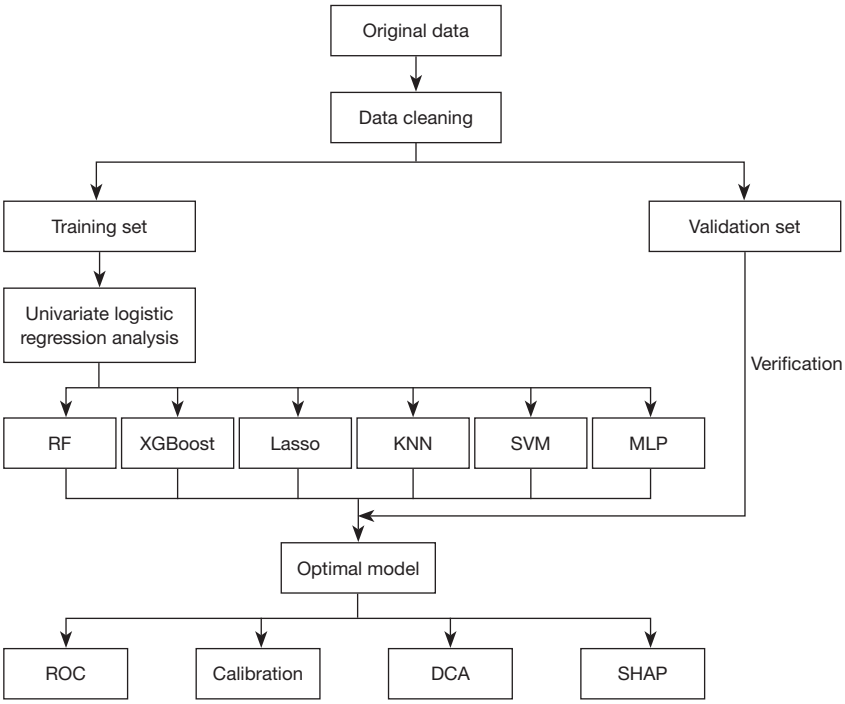

Figure 1 Technology roadmap. DCA, decision curve analysis; KNN, K-nearest-neighbors; Lasso, least absolute shrinkage and selection operator; MLP, multilayer perceptron; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine; SHAP, SHapley Additive exPlanations; XGBoost, extreme gradient boosting.

Table 1 Comparison between training set and validation set

Variables	Training set (n=228)	Validation set (n=99)	P
Coronary artery aneurysm			0.85
No	202 (88.6)	89 (89.9)	
Yes	26 (11.4)	10 (10.1)	
Sex			0.46
Male	82 (36.0)	40 (40.4)	
Female	146 (64.0)	59 (59.6)	
Incomplete Kawasaki disease			0.29
No	179 (78.5)	83 (83.8)	
Yes	49 (21.5)	16 (16.2)	
IVIG			0.73
No	8 (3.5)	2 (2.0)	
Yes	220 (96.5)	97 (98.0)	
Rash			0.67
No	55 (24.1)	21 (21.2)	
Yes	173 (75.9)	78 (78.8)	

Table 1 (*continued*)

Table 1 (*continued*)

Variables	Training set (n=228)	Validation set (n=99)	P
Conjunctival congestion			0.59
No	32 (14.0)	11 (11.1)	
Yes	196 (86.0)	88 (88.9)	
Swollen neck lymph node			0.35
No	59 (25.9)	31 (31.3)	
Yes	169 (74.1)	68 (68.7)	
Strawberry-like tongue			0.18
No	53 (23.2)	16 (16.2)	
Yes	175 (76.8)	83 (83.8)	
Rigid edema of hand and foot			0.71
No	140 (61.4)	58 (58.6)	
Yes	88 (38.6)	41 (41.4)	
Bacillus Calmette-Guérin scar			>0.99
No	217 (95.2)	95 (96.0)	
Yes	11 (4.8)	4 (4.0)	
Perianal change			0.44
No	190 (83.3)	79 (79.8)	
Yes	38 (16.7)	20 (20.2)	
Age (years)	2.00 (1.00, 4.00)	2.00 (1.00, 3.00)	0.61
Fever days before IVIG (days)	7.00 (6.00, 8.00)	7.00 (6.00, 7.00)	0.03
WBC (10^9/L)	15.12 (11.67, 19.44)	15.00 (11.70, 18.59)	0.74
NEU (10^9/L)	10.48 (7.08, 14.44)	10.23 (7.85, 13.32)	0.97
LYM (10^9/L)	3.01 (1.98, 4.51)	2.91 (2.11, 3.99)	0.80
MONO (10^9/L)	0.86 (0.61, 1.20)	0.89 (0.61, 1.12)	0.57
EOS (10^9/L)	0.19 (0.055, 0.50)	0.19 (0.04, 0.48)	0.79
BASO (10^9/L)	0.03 (0.02, 0.05)	0.03 (0.02, 0.04)	0.76
NR (%)	69.60 (57.40, 78.90)	70.6 (61.10, 78.10)	0.72
LR (%)	21.20 (13.20, 30.20)	22.0 (14.80, 30.20)	0.65
MR (%)	5.90 (4.10, 7.80)	5.90 (4.50, 7.30)	0.85
ER (%)	1.40 (0.40, 3.50)	1.40 (0.30, 3.40)	0.99
BR (%)	0.20 (0.10, 0.30)	0.20 (0.10, 0.30)	0.87
RBC (10^{12}/L)	4.03 (3.77, 4.33)	4.06 (3.77, 4.29)	0.74
Hb (g/L)	107.00 (100.00, 116.00)	108.00 (98.00, 115.00)	0.54
HCT (%)	32.5 (30.70, 35.20)	32.8 (31.00, 35.20)	0.83
MCV (fL)	82.2 (79.30, 84.50)	83.01 (79.60, 85.60)	0.12

Table 1 (*continued*)

Table 1 (*continued*)

Variables	Training set (n=228)	Validation set (n=99)	P
MCH (pg)	26.95 (25.95, 27.90)	27.00 (26.10, 28.20)	0.25
MCHC (g/L)	328.00 (322.00, 335.00)	328.00 (321.00, 335.00)	0.44
RDW-SD (fL)	39.90 (38.10, 41.70)	39.60 (38.20, 42.00)	0.74
RDW-CV (%)	13.30 (12.80, 13.90)	13.30 (12.80, 13.90)	0.72
PLT (10^9/L)	371.00 (309.00, 451.00)	365.00 (300.00, 444.00)	0.40
MPV (fL)	9.20 (8.60, 10.00)	9.3 (8.80, 10.20)	0.13
PCT (%)	0.35 (0.29, 0.43)	0.35 (0.28, 0.42)	0.77
PDW (%)	15.60 (15.40, 15.90)	15.7 (15.40, 15.90)	0.64
CRP (mg/L)	66.9 (36.00, 120.20)	73.8 (36.80, 115.30)	0.74
ESR (mm/h)	77 (56.00, 103.00)	75.00 (57.00, 98.00)	0.71
ALT (U/L)	26.50 (14.00, 90.00)	26.0 (11.00, 78.00)	0.32
AST (U/L)	36.00 (25.60, 59.00)	29.8 (23.00, 54.00)	0.19
TBil (μmol/L)	6.35 (4.30, 10.55)	7.00 (5.50, 10.70)	0.22
DBil (μmol/L)	2.50 (1.40, 4.40)	2.60 (1.70, 4.80)	0.25
CysC (mg/L)	0.80 (0.80, 1.00)	0.80 (0.80, 1.00)	0.59
TP (g/L)	68.40±7.50	68.60±6.80	0.79
ALB (g/L)	39.80 (37.30, 42.50)	41.1 (37.80, 44.70)	0.02
GGT (U/L)	32.00 (13.00, 109.00)	35.0 (10.90, 89.00)	0.27
PA (g/L)	80.00 (54.00, 98.00)	78.00 (57.00, 103.00)	0.71
AGR	1.50±0.30	1.50±0.30	0.08
Glb (g/L)	27.80 (24.40, 32.20)	27.4 (24.20, 30.70)	0.34
K^+ (mmol/L)	4.20±0.60	4.20±0.60	0.61
Na^+ (mmol/L)	136.60±3.00	136.60±3.80	0.81
Cl (mmol/L)	101.8 (99.00, 104.00)	101.9 (99.00, 104.00)	0.52

Data are presented as n (%), median (interquartile range), or mean ± standard deviation. ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALB, albumin; AGR, ratio of albumin to globulin; BASO, basophil; BR, basophilic ratio; CRP, C-reactive protein; CysC, cystatin C; DBil, direct bilirubin; EOS, eosinophilic; ER, eosinophilic ratio; ESR, erythrocyte sedimentation rate; GGT, gamma-glutamyl transferase; Glb, globulin; Hb, hemoglobin; HCT, hematocrit; IVIG, intravenous immunoglobulin; LYM, lymphocyte; LR, lymphocyte ratio; MONO, monocyte; MR, monocyte ratio; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular-hemoglobin concentration; MPV, mean platelet volume; NEU, neutrophil; NR, neutrophil ratio; PLT, platelet; PCT, plateletcrit; PDW, platelet distribution width; PA, prealbumin; RBC, red blood cell; RDW-SD, red cell distribution width-standard deviation; RDW-CV, red cell distribution width-coefficient of variation; TBil, total bilirubin; TP, total protein; WBC, white blood cell.

benefit (*Figure 5*).

Interpretation of feature contributions in XGBoost

For a given CAA positive sample, the sum of the SHAP values of all variables of XGBoost represents the individual deviation from the mean (E [f (x)] =–0.0516) predicted from the entire dataset (*Figure 6A*). The higher the overall SHAP value, the more the variable contributes to the CAA (*Figure 6B*). The SHAP total feature contribution plot provides a visual representation of the overall data (*Figure* 7). Each point represents a sample, with more yellow indicating higher

Table 2 Training set post Synthetic Minority Over-sampling Technique

Variable	Non-CAA (n=202)	CAA (n=196)	OR (95% CI)	P
Sex			0.754 (0.503, 1.129)	0.18
Male	72 (35.6)	83 (42.3)		
Female	130 (64.4)	113 (57.7)		
Incomplete Kawasaki disease			3.007 (1.939, 4.722)	<0.001
No	161 (79.7)	111 (56.6)		
Yes	41 (20.3)	85 (43.4)		
Rash			0.565 (0.360, 0.878)	0.01
No	45 (22.3)	66 (33.7)		
Yes	157 (77.7)	130 (66.3)		
Age (years)	2.00 (1.00, 4.00)	0.60 (0.30, 0.80)	0.093 (0.053, 0.150)	<0.001
Fever days before IVIG (days)	7.00 (6.00, 8.00)	7.00 (6.00, 10.00)	1.220 (1.108, 1.350)	<0.001
WBC (10^9/L)	15.14 (11.48, 19.54)	16.19 (13.24, 18.43)	0.999 (0.965, 1.035)	0.35
NEU (10^9/L)	10.59 (7.08, 14.57)	10.36 (8.06, 11.76)	0.957 (0.919, 0.996)	0.32
LYM (10^9/L)	2.87 (1.86, 4.28)	3.74 (3.06, 4.49)	1.327 (1.156, 1.533)	<0.001
MONO (10^9/L)	0.86 (0.61, 1.18)	1.29 (0.93, 1.51)	1.967 (1.450, 2.917)	<0.001
EOS (10^9/L)	0.18 (0.06, 0.46)	0.52 (0.28, 0.86)	2.759 (1.702, 4.616)	<0.001
BASO (10^9/L)	0.03 (0.02, 0.05)	0.04 (0.02, 0.05)	124,324.262 (414.400, 195,388,254.100)	<0.001
NR (%)	70.40 (57.80, 79.40)	62.70 (55.90, 66.60)	0.955 (0.939, 0.971)	<0.001
LR (%)	19.50 (12.60, 30.10)	23.30 (21.20, 27.20)	1.020 (1.001, 1.040)	0.002
MR (%)	5.80 (4.00, 7.60)	7.80 (6.00, 9.50)	1.197 (1.106, 1.300)	<0.001
ER (%)	1.30 (0.30, 3.10)	3.90 (2.10, 5.10)	1.314 (1.202, 1.444)	<0.001
BR (%)	0.20 (0.10, 0.30)	0.20 (0.10, 0.30)	0.648 (0.151, 2.659)	0.54
RBC (10^{12}/L)	4.08±0.45	3.56±0.58	0.135 (0.081, 0.215)	<0.001
Hb (g/L)	108.00 (101.00, 116.00)	93.00 (85.00, 102.00)	0.908 (0.888, 0.926)	<0.001
HCT (%)	33.00 (31.10, 35.50)	28.40 (25.40, 31.40)	0.747 (0.697, 0.796)	<0.001
MCV (fL)	82.30 (79.30, 84.50)	81.60 (77.20, 85.20)	1.011 (0.979, 1.046)	0.64
MCH (pg)	27.0 (26.10, 27.90)	26.50 (25.30, 27.80)	0.938 (0.843, 1.041)	0.04
MCHC (g/L)	328.00 (322.00, 335.00)	326.00 (322.00, 330.00)	0.962 (0.939, 0.985)	<0.001
RDW-SD (fL)	40.00 (38.10, 41.40)	38.80 (37.80, 41.70)	0.997 (0.946, 1.050)	0.07
RDW-CV (%)	13.30 (12.80, 13.90)	13.30 (12.90, 14.20)	1.015 (0.853, 1.211)	0.044
PLT (10^9/L)	362.00 (307.00, 449.00)	575.00 (473.00, 690.00)	1.010 (1.008, 1.012)	<0.001
MPV (fL)	9.20 (8.60, 10.00)	9.10 (8.50, 9.60)	0.737 (0.593, 0.912)	0.02
PCT (%)	0.34 (0.29, 0.42)	0.50 (0.41, 0.73)	8,069.605 (901.032, 93,398.233)	<0.001
PDW (%)	15.60 (15.40, 15.90)	15.30 (14.00, 15.60)	0.686 (0.581, 0.796)	<0.001
CRP (mg/L)	67.94 (36.65, 116.10)	50.28 (33.92, 69.38)	0.988 (0.983, 0.992)	<0.001
ESR (mm/h)	77.00 (56.00, 105.00)	69.00 (47.00, 84.00)	0.980 (0.972, 0.987)	<0.001

Table 2 (*continued*)

Table 2 (*continued*)

Variable	Non-CAA (n=202)	CAA (n=196)	OR (95% CI)	P
ALT (U/L)	26.00 (14.00, 100.00)	28.20 (20.50, 45.40)	0.992 (0.988, 0.996)	0.86
AST (U/L)	37.00 (26.00, 59.00)	32.50 (25.30, 40.20)	0.995 (0.990, 0.998)	0.003
TBil (μmol/L)	6.60 (4.30, 10.90)	7.90 (5.30, 11.10)	0.974 (0.949, 0.995)	0.049
DBil (μmol/L)	2.50 (1.40, 4.40)	4.10 (2.50, 6.50)	0.980 (0.949, 1.009)	<0.001
CysC (mg/L)	0.81 (0.74, 0.94)	0.96 (0.83, 1.10)	10.596 (3.626, 32.753)	<0.001
TP (g/L)	68.90±7.40	61.90±5.00	0.831 (0.795, 0.864)	<0.001
ALB (g/L)	40.20 (37.60, 43.10)	37.0 (35.80, 38.80)	0.732 (0.674, 0.790)	<0.001
GGT (U/L)	30.00 (13.00, 112.00)	54.1 (28.00, 113.90)	1.001 (0.998, 1.003)	<0.001
PA (g/L)	83.00 (60.00, 102.00)	50.00 (43.00, 59.00)	0.960 (0.951, 0.969)	<0.001
AGR	1.50±0.30	1.60±0.30	4.641 (2.418, 9.159)	<0.001
Glb (g/L)	28.10 (24.90, 32.20)	23.40 (20.30, 26.90)	0.826 (0.786, 0.866)	<0.001
K^+ (mmol/L)	4.18±0.61	4.50±0.46	3.052 (2.069, 4.612)	<0.001
Na^+ (mmol/L)	136.40 (134.10, 138.60)	136.20 (134.60, 138.50)	1.019 (0.953, 1.090)	0.84
Cl^- (mmol/L)	101.70 (99.00, 104.00)	102.30 (101.20, 103.50)	1.086 (1.022, 1.158)	0.04

Data are presented as n (%), median (interquartile range), or mean ± standard deviation, unless otherwise specified. ALT, alanine aminotransferase; AST, aspartate aminotransferase; ALB, albumin; AGR, ratio of albumin to globulin; BASO, basophil; BR, basophilic ratio; CAA, coronary artery aneurysms; CI, confidence interval; CRP, C-reactive protein; CysC, cystatin C; DBil, direct bilirubin; EOS, eosinophilic; ER, eosinophilic ratio; ESR, erythrocyte sedimentation rate; GGT, gamma-glutamyl transferase; Glb, globulin; Hb, hemoglobin; HCT, hematocrit; IVIG, intravenous immunoglobulin; LYM, lymphocyte; LR, lymphocyte ratio; MONO, monocyte; MR, monocyte ratio; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular-hemoglobin concentration; MPV, mean platelet volume; NEU, neutrophil; NR, neutrophil ratio; OR, odds ratio; PLT, platelet; PCT, plateletcrit; PDW, platelet distribution width; PA, prealbumin; RBC, red blood cell; RDW-SD, red cell distribution width-standard deviation; RDW-CV, red cell distribution width-coefficient of variation; TBil, total bilirubin; TP, total protein; WBC, white blood cell.

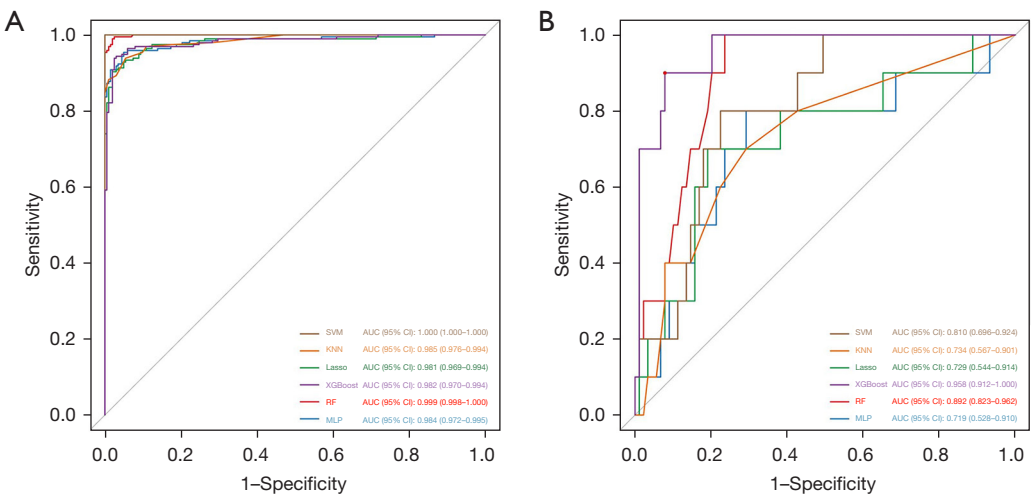

Figure 2 ROC curves of 6 machine learning models. (A) Training set; (B) validation set. AUC, area under the curve; KNN, K-nearest-neighbors; Lasso, least absolute shrinkage and selection operator; MLP, multilayer perceptron; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine; XGBoost, extreme gradient boosting.

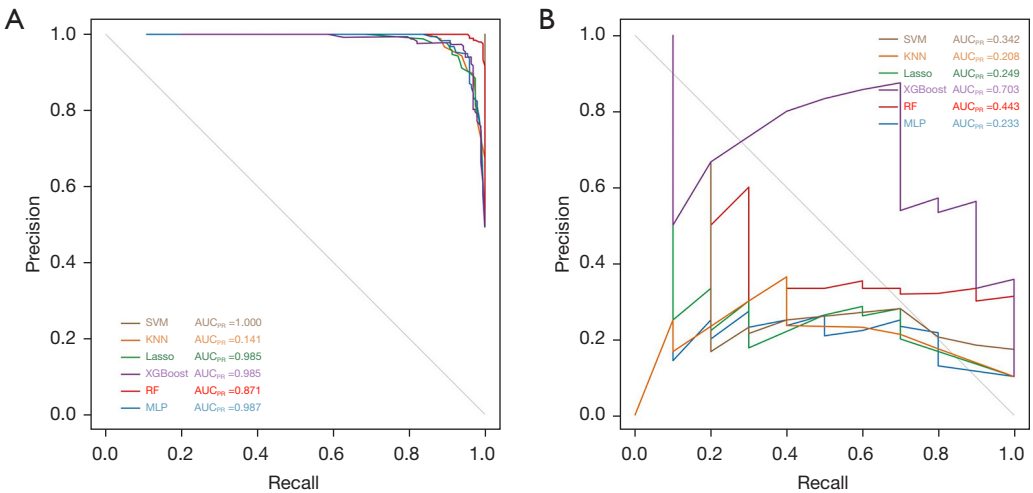

Figure 3 PR curves of 6 machine learning models. (A) Training set; (B) validation set. AUC, area under the curve; KNN, K-nearest-neighbors; Lasso, least absolute shrinkage and selection operator; MLP, multilayer perceptron; PR, Precision-Recall; RF, random forest; SVM, support vector machine; XGBoost, extreme gradient boosting.

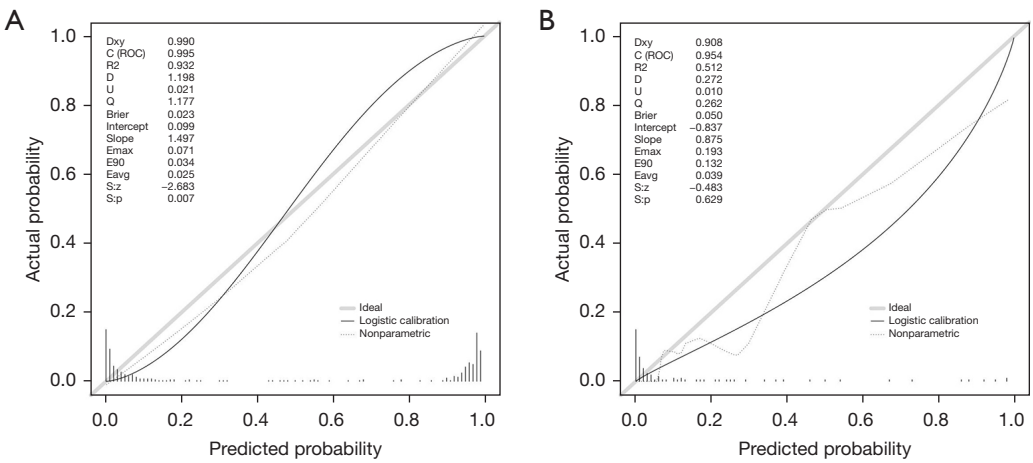

Figure 4 Calibration curves of XGBoost model in (A) training set and (B) validation set. C (ROC), area under the receiver operating characteristic curve; D, DeLong test; Dxy, Somers' Dxy Rank Correlation; E90, 90th percentile calibration error; Eavg, average calibration error; Emax, maximum calibration error; R^2, coefficient of determination; S:p, slope of the p-transformed calibration curve; S:z, slope of the z-transformed calibration curve; U, the concordance index (C-index); XGBoost, extreme gradient boosting.

feature values and more purple indicating lower feature values. When a feature's Shapley value is to the left of the center line, it indicates a negative value, resulting in a negative prediction by the model. Conversely, if the Shapley value is on the right side of the center line, it indicates a positive value, leading the model to a positive prediction.

Discussion

This study developed and validated an ML model using the XGBoost algorithm to predict CAA in children with KD, providing interpretable predictions that can improve clinical decision-making and guide personalized prevention

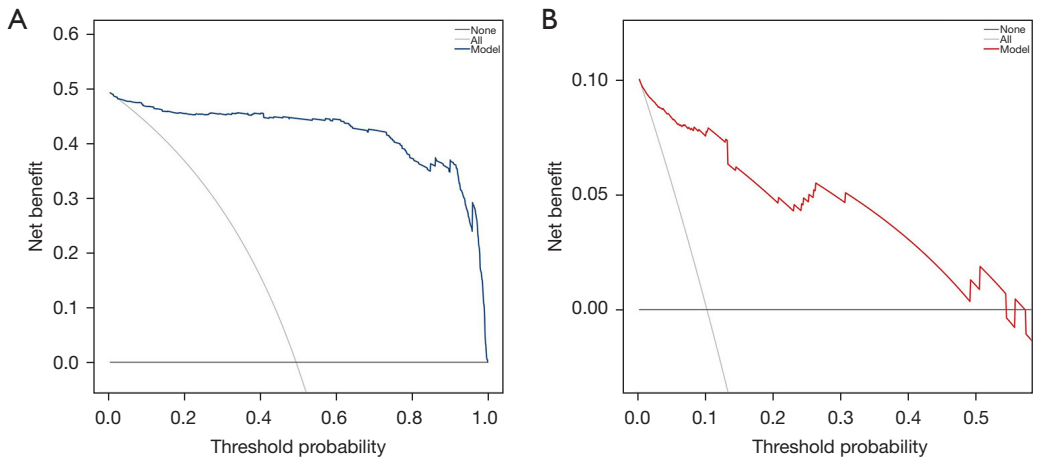

Figure 5 Decision curve analysis plot of XGBoost model in (A) training set and (B) validation set. XGBoost, extreme gradient boosting.

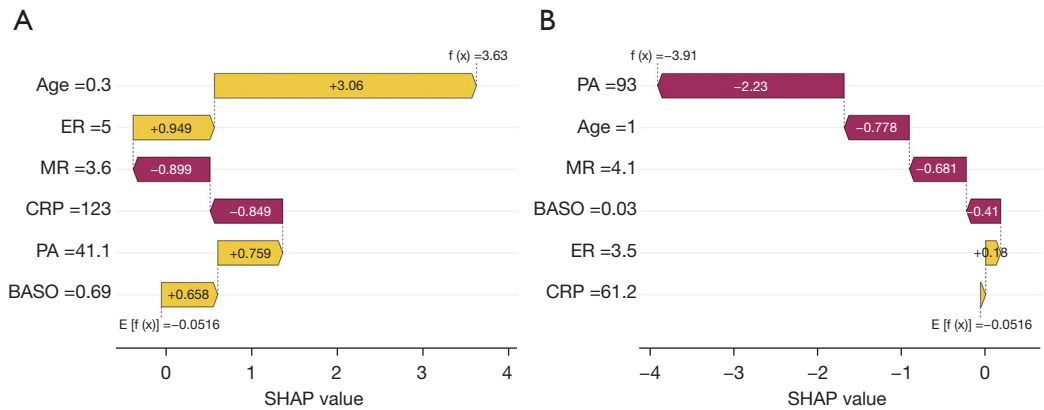

Figure 6 A given positive and negative single sample influence feature map. (A) Positive single sample; (B) negative single sample. BASO, basophil; CRP, C-reactive protein; ER, eosinophilic ratio; MR, monocyte ratio; PA, prealbumin; SHAP, SHapley Additive exPlanations.

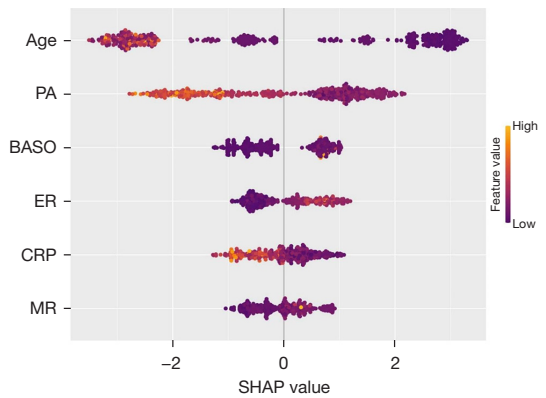

Figure 7 The overall feature map based on SHAP under the XGBoost model. BASO, basophil; CRP, C-reactive protein; ER, eosinophilic ratio; MR, monocyte ratio; PA, prealbumin; SHAP, SHapley Additive exPlanations; XGBoost, extreme gradient boosting.

strategies.

In this study, the XGBoost algorithm was considered optimal for predicting CAA, as it is an ensemble learning method based on gradient-boosted trees. This approach efficiently handles high-dimensional data and effectively mitigates overfitting. Its high accuracy and strong robustness make it an ideal choice for CAA prediction (15,16). Unlike logistic regression, which is a generalized linear model, in linear models, the weight of each feature can be interpreted as the importance of the feature, with the feature multiplied by its weight to determine the feature's impact on the prediction. Ensemble tree models, such as XGBoost, do not provide similar results as linear models, thus requiring the introduction of alternative methods to assess feature contribution. The SHAP model

is a comprehensive approach for model interpretability, offering both global explanations and the ability to assess the relationship between the model's predicted values and the included variables for individual samples, thereby explaining the model's predictions. The study covers a broad spectrum of clinical and laboratory variables, ensuring that the model captures the complexity of KD and its association with CAA. The inclusion of SHAP plots enhances the model's interpretability, crucial for clinician acceptance and integration into clinical practice. Multiple performance metrics, such as AUC, PR curve, calibration curve, and DCA, ensure the model's reliability and clinical utility. Additionally, the SMOTE addresses data imbalance, improving the model's detection of rare events like CAA. Finally, the inclusion of a validation set facilitates the assessment of the model's generalizability. When constructing ML models, the XGBoost package in R or the Python programming language with the scikit-learn and XGBoost libraries can be used for modeling and tuning.

Age is an important risk factor for the development of CAA in KD. A previous study showed that age is also an independent risk factor in prediction for moderate and giant CAA (17). Early-onset children (especially those younger than one year) have a higher risk of developing CAA (18), possibly due to atypical clinical presentations and differences in the maturity of the immune system and inflammatory responses (9). Conversely, older children have a lower, but not nonexistent, risk, emphasizing the importance of early diagnosis and treatment in all age groups.

In XGBoost algorithm-based predictive model, prealbumin (PA) is identified as a protective factor against CAA. This is consistent with the current clinical understanding of the relationship between PA and CAL (19). As an acute phase reactant, a decrease in PA during KD reflects increased inflammation in the body (20), which is closely related to the formation of CAA. Therefore, monitoring PA levels may be a valuable tool for the early detection of CAA risk in KD patients.

CRP, as an acute phase protein, reflects the inflammatory state in the body when elevated. Numerous previous studies have shown a positive correlation between elevated CRP levels and the occurrence of coronary artery disease (21,22). CRP levels have been found to be an independent risk factor for CAL in KD (23). However, the results of this study suggest that low CRP levels may also be associated with the occurrence of CALs. This phenomenon has been documented in case reports. Asai *et al.* (24) reported a case of a child with initially low CRP levels who eventually developed CAA. This discrepancy may be attributed to several factors. Firstly, as an acute-phase reactant, CRP levels are influenced not only by infection or inflammation but also by individual immune response variability, clinical presentation of the disease, and treatment protocols. Secondly, low CRP levels do not necessarily indicate the absence of an inflammatory response; other immune-mediated inflammatory factors, such as IL-6 and TNF-α, may also play significant roles in the development of CALs. Furthermore, there may be a time-dependent relationship between CRP elevation and CALs, where CRP levels measured early in the acute phase (e.g., within the first few days) may not fully reflect the entire inflammatory process, potentially leading to a lower CRP level despite the presence of CALs. Lastly, the development of CALs is a complex multifactorial process, influenced not only by inflammation but also by genetic factors, vascular wall damage, and hemodynamic factors. Therefore, in certain patient populations, low CRP levels should not exclude the possibility of CALs.

Leukocyte subtypes, such as eosinophils, monocytes, and basophils, have been identified as factors contributing to the development of CAL. Studies have shown that monocytes and other cells can lead to abnormal cytokine production following viral infections. These pro-inflammatory cytokines and chemokines play a critical role in the induction and regulation of various immune responses (25). The interplay between inflammation and oxidative stress may further amplify each other, potentially triggering the onset of KD.

In this model, absolute basophil count is identified as a risk factor for CAA in KD. As part of the immune system, basophils are involved in the regulation of inflammation and immune responses. Upon activation, basophils release a variety of active mediators (histamines, leukotrienes, cytokines, etc.) that mediate tissue damage (26). Traditionally, basophils were thought to play an auxiliary role in allergic inflammation, assisting mast cells in releasing inflammatory mediators. However, recent studies have demonstrated the critical role of basophils in initiating and sustaining allergic immune responses. For example, it is found that basophils were highly activated and significantly reduced in the peripheral blood of patients with rheumatoid arthritis, possibly migrating to secondary lymphoid tissues and sites of inflammation to counteract the peripheral Th1 response imbalance (27). It is reported that basophils have a reference value in clinical screening and diagnosis of allergic inflammatory diseases (28). In the context of KD,

an increase in basophil count may indicate an immune regulatory change, possibly related to CAA formation. The predictive value of this finding requires further research.

This study shows that an increase in eosinophils is an independent risk factor for the development of CAA. Eosinophils play an active role in allergic diseases as they are key players in immune and allergic responses. It has been found that KD patients with lower Th2 immune responses (such as lower levels of IL-5 or eosinophils) have a higher risk of CAL, suggesting that Th2 responses may play a protective role in the development of CALs in KD (29). It's found that eosinophil levels were significantly elevated in KD patients, and early combined detection is of great importance in prediction for KD and its associated CAL (30). A previous report has shown that eosinophils play a crucial role in nomogram models and are one of the most important predictors of KD (31).

Monocytes, as circulating WBC, are important in both innate and adaptive immunity, playing a major role in immune defense, inflammation, and tissue remodeling (32). Activated platelets interact with various WBC, including monocytes and neutrophils, triggering intercellular signaling leading to thrombosis and synthesis of inflammatory mediators. This may be involved in the inflammation and development of CALs in KD.

There are several limitations that should be considered when interpreting our study results. First, this study employed a retrospective design, which may lead to data bias and insufficient inference of causal relationships. As retrospective studies cannot control for all potential confounding factors, we are unable to exclude the possibility of unobserved variables influencing the outcomes. Second, the sample size is relatively small, particularly with a limited number of patients with CAL, which restricts our ability to fully assess the impact of various clinical features or therapeutic interventions on inflammatory biomarker levels. The small sample size may reduce the precision of statistical analyses and the generalizability of the results. Therefore, future studies should involve larger patient cohorts to validate our findings. Third, the use of oversampling techniques in this study may have influenced the representativeness of the results. There may have been over-selection of certain patient groups (e.g., those with more severe CAA), making the findings less representative of the overall patient population. Hence, future research should consider more balanced sample selection to enhance the external validity of the results.

Conclusions

In conclusion, this study developed and validated an explainable model for prediction of CAA in KD. The XGBoost algorithm was used to develop an explainable prediction model for CAA in KD based on age, prealbumin, absolute basophil count, eosinophil percentage, CRP, and monocyte percentage. These factors provide interpretable prediction results that enable physicians to understand the model's decision-making process, increase their confidence in the model, and effectively use ML model predictions in clinical decision making to provide more accurate and personalized prevention and treatment plans for patients.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tp.amegroups.com/article/view/10.21037/tp-24-359/rc

Data Sharing Statement: Available at https://tp.amegroups.com/article/view/10.21037/tp-24-359/dss

Peer Review File: Available at https://tp.amegroups.com/article/view/10.21037/tp-24-359/prf

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tp.amegroups.com/article/view/10.21037/tp-24-359/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Medical Ethics Committee of Nanchong Central Hospital [No. 2024(024)]. This article is a retrospective study. Therefore, the Medical Ethics Committee of Nanchong Central Hospital waived the requirement to obtain distinct written informed consent from the patients.

References

1. Butters C, Curtis N, Burgner DP. Kawasaki disease fact check: Myths, misconceptions and mysteries. J Paediatr Child Health 2020;56:1343-5.
2. Soni PR, Noval Rivas M, Arditi M. A Comprehensive Update on Kawasaki Disease Vasculitis and Myocarditis. Curr Rheumatol Rep 2020;22:6.
3. Leung PB, Davis AM, Kumar S. Diagnosis and Management of Nonalcoholic Fatty Liver Disease. JAMA 2023;330:1687-8.
4. Ae R, Abrams JY, Maddox RA, et al. Outcomes in Kawasaki disease patients with coronary artery abnormalities at admission. Am Heart J 2020;225:120-8.
5. Bang JS, Kim GB, Kwon BS, et al. Long-Term Prognosis for Patients with Kawasaki Disease Complicated by Large Coronary Aneurysm (diameter ≥6 mm). Korean Circ J 2017;47:516-22.
6. Gong X, Tang L, Wu M, et al. Development of a nomogram prediction model for early identification of persistent coronary artery aneurysms in kawasaki disease. BMC Pediatr 2023;23:79.
7. Liu J, Yue Q, Qin S, et al. Risk factors and coronary artery outcomes of coronary artery aneurysms differing in size and emergence time in children with Kawasaki disease. Front Cardiovasc Med 2022;9:969495.
8. Tang Y, Ding C, Xu Q, et al. Prediction nomogram for coronary artery aneurysms at one month in Kawasaki disease. Ital J Pediatr 2023;49:146.
9. Son MBF, Gauvreau K, Tremoulet AH, et al. Risk Model Development and Validation for Prediction of Coronary Artery Aneurysms in Kawasaki Disease in a North American Population. J Am Heart Assoc 2019;8:e011319.
10. Miyata K, Miura M, Kaneko T, et al. Evaluation of a Kawasaki Disease Risk Model for Predicting Coronary Artery Aneurysms in a Japanese Population: An Analysis of Post RAISE. J Pediatr 2021;237:96-101.e3.
11. Tang Y, Liu Y, Du Z, et al. Prediction of coronary artery lesions in children with Kawasaki syndrome based on machine learning. BMC Pediatr 2024;24:158.
12. Zhang Y, Liu P, Tang LJ, et al. Basing on the machine learning model to analyse the coronary calcification score and the coronary flow reserve score to evaluate the degree of coronary artery stenosis. Comput Biol Med 2023;163:107130.
13. Sharma A, Lysenko A, Boroevich KA, et al. DeepFeature: feature selection in nonimage data using convolutional neural network. Brief Bioinform 2021;22:bbab297.
14. Newburger JW, Takahashi M, Gerber MA, et al. Diagnosis, treatment, and long-term management of Kawasaki disease: a statement for health professionals from the Committee on Rheumatic Fever, Endocarditis and Kawasaki Disease, Council on Cardiovascular Disease in the Young, American Heart Association. Circulation 2004;110:2747-71.
15. Heo J, Yoo J, Lee H, et al. Prediction of Hidden Coronary Artery Disease Using Machine Learning in Patients With Acute Ischemic Stroke. Neurology 2022;99:e55-65.
16. Sunaga Y, Watanabe A, Katsumata N, et al. A simple scoring model based on machine learning predicts intravenous immunoglobulin resistance in Kawasaki disease. Clin Rheumatol 2023;42:1351-61.
17. Jiang S, Li M, Xu K, et al. Predictive factors of medium-giant coronary artery aneurysms in Kawasaki disease. Pediatr Res 2024;95:267-74.
18. Fabi M, Andreozzi L, Frabboni I, et al. Non-coronary cardiac events, younger age, and IVIG unresponsiveness increase the risk for coronary aneurysms in Italian children with Kawasaki disease. Clin Rheumatol 2021;40:1507-14.
19. Zhu Y, Yu Z, Xu R, et al. Associations of serum high-sensitivity C-reactive protein and prealbumin with coronary vessels stenosis determined by coronary angiography and heart failure in patients with myocardial infarction. J Med Biochem 2023;42:9-15.
20. Huang YT, Jiang MY, Hwang JC. Albumin to prealbumin ratio in peritoneal dialysis patients: Clinical implication and outcome prediction. PLoS One 2022;17:e0276159.
21. He L, Fan C, Li G. The relationship between serum C-reactive protein and senile hypertension. BMC Cardiovasc Disord 2022;22:500.
22. Bisaria S, Terrigno V, Hunter K, et al. Association of Elevated Levels of Inflammatory Marker High-Sensitivity C-Reactive Protein and Hypertension. J Prim Care Community Health 2020;11:2150132720984426.
23. Shuai S, Zhang H, Zhang R, et al. Prediction of coronary artery lesions based on C-reactive protein levels in children

with Kawasaki Disease: a retrospective cohort study. J Pediatr (Rio J) 2023;99:406-12.
24. Asai Y, Suzuki T, Yamada M. An infant case of coronary artery aneurysms with no systemic symptoms after treatment for Kawasaki disease. Cardiol Young 2024. [Epub ahead of print]. doi: 10.1017/S1047951124025563.
25. Philip S, Jindal A, Krishna Kumar R. An update on understanding the pathophysiology in Kawasaki disease: Possible role of immune complexes in coronary artery lesion revisited. Int J Rheum Dis 2023;26:1453-63.
26. Kahlenberg JM, Kang I. Advances in Disease Mechanisms and Translational Technologies: Clinicopathologic Significance of Inflammasome Activation in Autoimmune Diseases. Arthritis Rheumatol 2020;72:386-95.
27. Koppejan H, Hameetman M, Beyrend G, et al. Immunoprofiling of early, untreated rheumatoid arthritis using mass cytometry reveals an activated basophil subset inversely linked to ACPA status. Arthritis Res Ther 2021;23:272.
28. Chang LS, Huang YH, Chang HY, et al. Basophils Predict Mite Sensitization in Patients with Kawasaki Disease. Children (Basel) 2023;10:1209.
29. Chang L, Yang HW, Lin TY, et al. Perspective of Immunopathogenesis and Immunotherapies for Kawasaki Disease. Front Pediatr 2021;9:697632.
30. Chang LS, Chen KD, Huang YH, et al. Expression of Eosinophilic Subtype Markers in Patients with Kawasaki Disease. Int J Mol Sci 2022;23:10093.
31. Liu XP, Huang YS, Xia HB, et al. A Nomogram Model Identifies Eosinophilic Frequencies to Powerfully Discriminate Kawasaki Disease From Febrile Infections. Front Pediatr 2020;8:559389.
32. Kratofil RM, Kubes P, Deniset JF. Monocyte Conversion During Inflammation and Injury. Arterioscler Thromb Vasc Biol 2017;37:35-42.

Cite this article as: Song Z, Ming H, Liu B, Liu D. Development and validation of an explainable machine learning-based prediction model for primary Kawasaki disease complicated with coronary artery aneurysms. Transl Pediatr 2025;14(2):208-221. doi: 10.21037/tp-24-359