scientific reports

Check for updates

OPEN

Predicting and interpreting key features of refractory *Mycoplasma pneumoniae* pneumonia using multiple machine learning methods

Yuhan Jiang[1,2,5,6], Xu Wang[3,6], Li Li[4,6], Yifan Wang[1,2], Xuelin Wang[1] & Yingxue Zou[1,2,5]✉

In recent years, the incidence of refractory *Mycoplasma pneumoniae* pneumonia (RMPP) has significantly risen, posing severe pulmonary and extrapulmonary complications, making early identification a challenge for clinicians. In this retrospective single-center study, we included patients diagnosed with *Mycoplasma pneumoniae* pneumonia in 2021, categorizing them into RMPP and non-RMPP groups. Univariate regression analysis initially identified variables associated with RMPP. Seven mainstream machine learning methods were then employed to construct predictive models, evaluated for reliability and robustness through tenfold cross-validation and sensitivity analysis. Ultimately, the optimal predictive model was selected using multidimensional metric assessments, and SHAP analysis identified key predictive factors related to RMPP. Twenty-nine factors from various dimensions were found to be associated with RMPP and used to build the predictive model. The XGBoost model demonstrated high predictive capability with an accuracy of 0.80 and an AUC of 0.93. Ten-fold cross-validation and sensitivity analysis confirmed the model's robustness and reliability. SHAP analysis interpreted the final model with 8 key features. These features include fever duration, macrolide treatment before hospitalization, severe *Mycoplasma pneumoniae* pneumonia, lactate dehydrogenase, neutrophil-to-lymphocyte ratio, alanine aminotransferase, peak fever, and extensive lung consolidation. This simple, effective predictive model enhances clinicians' understanding and aids early identification of RMPP.

Keywords *Mycoplasma Pneumoniae*, Refractory *Mycoplasma Pneumoniae* pneumonia, Machine learning, Predictive model

Abbreviations

RMPP	Refractory *Mycoplasma pneumoniae* Pneumonia
MPP	*Mycoplasma pneumoniae* Pneumonia
MP	*Mycoplasma pneumoniae*
SHAP	SHapley Additive exPlanations
Ig	Immunoglobulin
ALT	Alanine Aminotransferase
CRP	C-reactive protein
SMPP	Severe *Mycoplasma pneumoniae* pneumonia
NLR	Neutrophil-to-lymphocyte ratio
PLR	Platelet-to-lymphocyte ratio
MPVLR	Mean platelet volume-to-lymphocyte ratio
RF	Random Forest
XGBoost	EXtreme gradient boosting
NB	Naive Bayes
SVM	Support vector machine

[1]Tianjin Children's Hospital (Children's Hospital, Tianin University), Tianjin, China. [2]Clinical School of Pediatrics, Tianjin Medical University, Tianjin, China. [3]Tianjin Women and Children's Health Center, 96 Guizhou Road, Heping District, Tianjin, China. [4]Department of Pediatrics, Second Affiliated Hospital of Guangxi Medical University, Guangxi, China. [5]Tianjin Key Laboratory of Birth Defects for Prevention and Treatment, Tianjin, China. [6]Yuhan Jiang, Xu Wang and Li Li have contributed equally to this work. ✉email: zouyingxue2015@tju.edu.cn

DT	Decision tree
KNN	K-Nearest neighbors
ROC	Receiver operating characteristic
AUC	Area under the ROC curve
LDH	Lactate dehydrogenase
CD	Cluster of differentiation
CK	Creatine kinase
Fer	Ferritin

Mycoplasma pneumoniae (MP) is one of the most common respiratory pathogens among school-aged children, with infections following an epidemic cycle that occurs every 3 to 5 years, lasting 1 to 2 years each time[1,2]. Typically presenting mild symptoms, *Mycoplasma pneumoniae* pneumonia (MPP) often manifests as a self-limiting disease[3,4]. Despite this, there are still some children with MPP who remain febrile, with worsening clinical symptoms and radiological findings or develop extrapulmonary complications, after receiving standard treatment with macrolide antibiotics for 7 days or more. This condition is defined as refractory *Mycoplasma pneumoniae* pneumonia (RMPP)[5,6]. The escalating challenge of antibiotic treatment, particularly in the context of growing antibiotic resistance, has led to a significant rise in the incidence of RMPP[7,8], prompting considerable attention and further exploration.

Despite treatment with effective antimicrobial drugs for MPP, alterations in the host's immune response may lead to a severe inflammatory process, causing damage[9]. Some patients may even face serious complications such as necrotizing pneumonia, pulmonary atelectasis, pulmonary embolism, empyema, and in extreme cases, life-threatening situations[10–13]. The early identification of RMPP has historically been challenging for clinicians. Therefore, the quest for reliable, sensitive, and early identification methods has become crucial to better understand and manage RMPP.

While prior research has made strides in exploring potential predictive factors for RMPP[14], these conventional statistical methods have shown limitations in predictive efficacy[3], such as limited ability to handle non-linear relationships and interactions, as well as inadequate adaptability to complex data structures[15,16]. In this context, machine learning emerges as a powerful tool capable of handling large volumes of data and uncovering complex, latent relationships between variables[17]. Currently, machine learning approaches have demonstrated notable efficacy in predicting MPP based on routine blood parameters and imaging, outperforming traditional regression methods in other areas such as disease detection, risk prediction, and treatment guidance[18–20]. Furthermore, the intricate and often incomprehensible predictive factors computed by machine learning may not directly guide clinical practice. SHAP (SHapley Additive exPlanations)[21], however, can elucidate the outputs of machine learning models[22], identifying features that exert the most significant influence on model outputs[23], thereby enhancing our understanding of the contributions and impacts of these factors on disease progression.

Hence, this study aims to utilize machine learning methodologies to select an optimal predictive model for RMPP, explore relevant predictive factors, and employ SHAP ranking to elucidate the importance of these factors, intending to provide critical guidance and decision support for future clinical practices.

Methods

Study setting and participants

This study conducted a retrospective analysis involving 1335 hospitalized pediatric patients diagnosed with MPP at Tianjin Children's Hospital between January and December 2021. The sample size possesses an appropriate effect size and machine learning accuracy (≥ 80%)[24].

Inclusion criteria comprised[1]: patients aged greater than 1 month and less than 18 years old,[2] diagnosis of MPP as defined by the "Expert Consensus on Diagnosis and Treatment of Childhood Mycoplasma Pneumoniae Pneumonia (2023 Edition)[6,25]" as follows: In addition to clinical symptoms and signs of pneumonia, patients must have a positive result from one of the following laboratory tests:[1] Detection of MP-DNA or MP-RNA by nucleic acid amplification test. (2) Single-test titers of anti-MP antibodies ≥ 1:160, or a fourfold increase in antibody titers during the course of illness by MP antibody serum agglutination test. Exclusion criteria included: (1) a history of recurrent respiratory tract infections, defined as "≥ 8 episodes per year in children under 3 years old and ≥ 6 episodes per year in children over 3 years old." These recurrent infections are often associated with underlying primary diseases, which may introduce potential confounding effects, and (2) comorbidities such as severe illnesses or congenital immunodeficiency[26,27].

Variable selection

The variables were retrospectively extracted from the hospital's electronic medical records system and systematically classified into three distinct groups: (1) Clinical data comprising demographic information, medical history, comorbidities, and clinical symptoms; (2) Laboratory test results encompassing complete blood count (white blood cell count, neutrophils, lymphocytes, hemoglobin, platelets, etc.), biochemical parameters (blood glucose, sodium, potassium, calcium, etc.), liver and kidney function tests (alanine aminotransferase (ALT), aspartate aminotransferase, total bilirubin, albumin, etc.), immune function tests (immunoglobulin (Ig) levels, C-reactive protein (CRP), etc.), and coagulation indicators (prothrombin time, activated partial thromboplastin time, fibrinogen, D-dimer, etc.). Additionally, all laboratory data were measured after hospitalization; and (3) Radiological features including lung collapse, pleural effusion, and lobar consolidation. Additionally, microbiological data were collected to identify potential pathogens, including sputum, throat, and nasal swabs for respiratory pathogens. To enhance statistical efficiency, and maintain focus on the comprehensive prediction of RMPP, the detected pathogens were categorized into a binary variable, "co-infection" to evaluate their potential impact on study outcomes.

In addition, we generated several composite indicators, such as the neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), mean platelet volume-to-lymphocyte ratio (MPVLR), and severe *Mycoplasma pneumoniae* pneumonia (SMPP), which is diagnosed as confirmed MPP and additionally meets any of the following criteria(6): (1)poor general condition; (2)altered consciousness, cyanosis, or tachypnea; (3)requirement for assisted ventilation or intermittent apnea with oxygen saturation ≤ 92%; (4)hyperpyrexia lasting more than 5 days; (5)dehydration and refusal of oral intake; (6)radiographic evidence of ≥ 2/3 unilateral lung consolidation, multilobar infiltration, pleural effusion, pneumothorax, atelectasis, lung necrosis, and lung abscess; and (7) extrapulmonary complications. Detailed criteria are presented in Table S1. MP resistance testing targets sites 2063 and 2064 on the central loop of the fifth structure domain (V region) of the 23S rRNA, with the testing specimen being bronchoalveolar lavage fluid.

Statistical analysis

Based on the definition of RMPP, patients are categorized into RMPP and non-RMPP groups. To maintain data analytical reliability, variables with over 20% missing data or fewer than 80 cases in binary classification were excluded. Subsequently, univariate logistic regression analysis identified variables linked to RMPP using the entire dataset. Given the assumption that missingness is random, multiple imputations were employed to address missing values in these variables. The Spearman correlation coefficients were subsequently used to control for collinearity among the variables[16,28].

After standardizing the continuous variables, the selected variables were integrated into the model. Before performing the machine learning analysis, we randomly split the samples into training and testing sets in a 3:1 ratio, ensuring that there was no overlap between the two sets. To comprehensively assess their performance and robustness across different algorithms, seven distinct machine learning models: logistic regression, random forest (RF), eXtreme Gradient Boosting (XGBoost), naive Bayes (NB), support vector machine (SVM), decision tree (DT), and K-nearest neighbors (KNN) —were employed to predict outcome probabilities in this study. This approach aimed to identify the most suitable model for our dataset. To mitigate class imbalance within RMPP, undersampling techniques were implemented during model training[29]. Moreover, tenfold cross-validation was executed to ensure robust and dependable outcomes. Model performance based on metrics like accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC), were conducted to select the optimal predictive model from the data. It is important to note that all model training, tenfold cross-validation, and SHAP-based analyses were conducted solely on the training set. Model performance evaluation was carried out exclusively on the independent testing set to ensure objectivity and assess generalizability.

Sensitivity analysis and variable importance

To explore the role of the MP resistance test in the occurrence and progression of RMPP, a sensitivity analysis was conducted on a subset of individuals who underwent resistance gene site testing within the best predictive model. Additionally, to interpret the outputs of the machine learning model, SHAP feature importance and permutation importance were employed to clarify the relationship between the optimal predictive model and its key predictors. Decision thresholds for the corresponding variables were visualized to facilitate readers' understanding. Finally, dependence plots were utilized to explore the nonlinear impact of individual features on the entire dataset. All statistical analyses were performed using Python (version 3.8.2), with a two-sided P value of < 0.05 considered statistically significant.

Results

Population and patient characteristics

A total of 1,332 patients were included in the final analysis. Based on the RMPP criteria, patients were categorized into two groups: 1,210 (90.8%) in the non-RMPP group and 122 (9.2%) in the RMPP group. Furthermore, patients were classified according to the SMPP criteria, with 986 (74.1%) in the non-SMPP group and 346 (25.9%) in the SMPP group. Among the 346 SMPP patients, 262 (75.7%) were in the non-RMPP group, while 84 (24.3%) were in the RMPP group. The demographic characteristics are shown in Tables 1 and 2. The initial dataset contained 175 variables. Univariate analyses between the two groups revealed statistically significant differences in RMPP related to the following clinical features: age, duration of fever days, peak fever, glucocorticoid treatment before hospitalization, received macrolide treatment before hospitalization, respiratory rate, reduced respiratory sounds, wheezing, rash, abdominal pain, and SMPP. Significant differences in laboratory serum indices were observed for lymphocytes, CRP, serum calcium, sodium, creatine kinase (CK), lactate dehydrogenase (LDH), ALT, IgM, IgA, ferritin (Fer), cluster of differentiation (CD)3, CD4, NLR, MPVLR, and PLR. Radiological features showing statistical differences included pleural thickening, pleural effusion, lung collapse, and extensive lung consolidation. Through Spearman correlation coefficient calculations, the CD4 variable with linear correlation (correlation coefficient > 0.8) was excluded. Consequently, 29 variables were selected after multiple imputations for integration into the predictive model. The detailed results are shown in Table 2.

Machine learning model evaluation

After randomly splitting the dataset into training and testing sets at a 3:1 ratio, we examined the differences between the two subsets in terms of evaluation variables. Most variables exhibited no statistically significant differences between the training and testing sets, suggesting an adequately balanced data split as detailed in Table S2. Figure 1A displays the ROC curve of the original data, based on evaluation using the test dataset. Table 3 showcases the performance metrics of each model during evaluation. The XGBoost model showed a high level of accuracy (0.80) and AUC (0.93), maintaining a recall (0.97). The logistic regression model performed well in accuracy (0.85) and AUC (0.89) but showed a slightly lower recall (0.74). The KNN model achieved a balanced

	Non-RMPP	RMPP	Overall
	(*N* = 1210)	(*N* = 122)	(*N* = 1332)
Sex			
Male	595 (49.2%)	69 (56.6%)	664 (49.8%)
Female	615 (50.8%)	53 (43.4%)	668 (50.2%)
Age			
Mean (SD)	5.21 (3.01)	6.28 (2.92)	5.31 (3.02)
Median [Min, Max]	4.00 [1.00, 16.0]	6.00 [1.00, 16.0]	5.00 [1.00, 16.0]
Season of hospitalization			
Spring	83 (6.9%)	6 (4.9%)	89 (6.7%)
Summer	149 (12.3%)	7 (5.7%)	156 (11.7%)
Autumn	222 (18.3%)	23 (18.9%)	245 (18.4%)
Winter	756 (62.5%)	86 (70.5%)	842 (63.2%)
Duration of hospitalization			
Mean (SD)	5.95 (1.87)	8.98 (3.17)	6.23 (2.20)
Median [Min, Max]	6.00 [2.00, 15.0]	9.00 [3.00, 21.0]	6.00 [2.00, 21.0]
Oxygen therapy			
High-flow oxygen	7 (0.6%)	1 (0.8%)	8 (0.6%)
Low-flow oxygen	1203 (99.4%)	121 (99.2%)	1324 (99.4%)
Invasive mechanical ventilation			
Yes	0 (0%)	0 (0%)	0 (0%)
No	1210 (100%)	122 (100%)	1332 (100%)
Extrapulmonary complications			
Yes	74 (6.1%)	12 (9.8%)	86 (6.5%)
No	1136 (93.9%)	110 (90.2%)	1246 (93.5%)

Table 1. General Characteristics of RMPP and Non-RMPP Data. *RMPP: refractory *Mycoplasma pneumoniae* pneumonia.

result across metrics, maintaining relative stability in accuracy (0.85) and AUC (0.90). The RF model stood out in accuracy (0.85) and AUC (0.93), exhibiting a slight increase in recall (0.89) compared to other models. The SVM model showed moderate performance in accuracy (0.80) and AUC (0.85). NB and DT models performed at an intermediate level, with the NB model's accuracy (0.76) slightly lower than the DT model (0.78).

Figure 1B shows the ROC curves derived from tenfold cross-validation with the training dataset, with corresponding model performance metrics summarized in Table 3. Compared to their performance on the original data, most models exhibited a slight decrease in performance after cross-validation. Notably, the XGBoost model displayed improved accuracy (0.91), and precision (0.50), indicating enhancement in both stability and accuracy of the model. These improved metrics suggest that the XGBoost model may offer higher reliability and stability in predicting RMPP, demonstrating more pronounced comprehensive performance.

Sensitivity analysis of MP resistance

Multiple pieces of evidence suggest the importance of a positive *Mycoplasma* resistance test as a predictive indicator for RMPP. However, due to a missing rate exceeding 20% in previous models' data quality control, this study excluded this variable. Subsequently, a sensitivity analysis was conducted using a small subset of 292 individuals who underwent Mycoplasma resistance testing. Among these individuals, 263 (90%) tested positive for resistance. In the RMPP group, the resistance proportion was approximately 96.5% (55/57), while in the non-RMPP group, the resistance proportion was 88.5% (208/235). Univariate model analysis revealed no statistically significant correlation between *Mycoplasma* resistance and RMPP ($P > 0.05$). Two XGBoost predictive models were compared in our analysis: Model 1 included the MP resistance test variable and 29 related predictive factors, while Model 2 comprised only the 29 associated predictive factors, excluding the MP resistance gene test variable. Model 1 exhibited slightly higher precision and recall than Model 2. However, Model 2 displayed a slightly higher AUC than Model 1 (0.851 vs. 0.846), suggesting that including this variable did not notably improve the model's predictive capacity. Furthermore, SHAP variable sorting in Model 1 revealed a relatively lower importance ranking of *Mycoplasma* resistance detection for predicting RMPP, detailed in Table S3 and Fig. 2A.

Variable importance analysis of RMPP

The prominent predictive model, XGBoost, was chosen to determine variable importance using SHAP analysis based on the training dataset. Figure 2B displays the results, ranking features by weightage using XGboost. The top 8 variables were identified as the strongest predictors of RMPP. A simple predictive model was constructed using these variables, demonstrating high accuracy (0.91) and AUC (0.90). These few simple variables were used as the final model for application development. The 8 variables include fever days and external macrolide

Variable	RMPP		OR [CI5, CI95]	Missing rate	P
	No ($n=1210$)	Yes ($n=122$)			
Age (mean (SD))	5.21 (3.01)	6.28 (2.92)	1.1147 [1.0522, 1.1810]	0.00%	< 0.001
Duration of Fever days (mean (SD))	5.72 (3.00)	10.05 (2.91)	1.6339 [1.5066, 1.7721]	0.00%	< 0.001
Peak Fever (mean (SD))	39.10 (1.41)	39.68 (1.10)	2.4418 [1.8656, 3.1960]	0.45%	< 0.001
Macrolide Treatment (%)	500 (41.3)	103 (84.4)	7.6979 [4.6570,12.7244]	0.00%	< 0.001
Glucocorticoid Treatment (%)	85 (7.0)	29 (23.8)	4.1271 [2.5754, 6.6138]	0.00%	< 0.001
Respiratory Rate (mean (SD))	30.17 (4.77)	31.77 (4.99)	1.0581 [1.0238, 1.0935]	0.00%	< 0.001
Reduced Respiratory Sounds (%)	118 (9.8)	32 (26.2)	3.2904 [2.1067, 5.1391]	0.00%	< 0.001
Wheezing (%)	257 (21.2)	13 (10.7)	0.4423 [0.2448, 0.7989]	0.00%	< 0.01
Rash (%)	61 (5.0)	19 (15.6)	3.4746 [1.9987, 6.0405]	0.00%	< 0.001
Abdominal Pain (%)	81 (6.7)	17 (13.9)	2.2567 [1.2893, 3.9500]	0.00%	< 0.01
SMPP (%)	262 (21.7)	84 (68.9)	7.9984 [5.3248,12.0145]	0.00%	< 0.001
Pleural Effusion (%)	61 (5.0)	35 (28.7)	7.5777 [4.7394,12.1158]	0.00%	< 0.001
Pleural Thickening (%)	440 (36.4)	71 (58.2)	2.4363 [1.6688, 3.5567]	0.00%	< 0.001
Lung Collapse (%)	111 (9.2)	28 (23.0)	2.9492 [1.8526, 4.6949]	0.00%	< 0.001
Extensive Lung Consolidation (%)	222 (18.3)	64 (52.5)	4.9108 [3.3450, 7.2098]	0.00%	< 0.001
L(10^9/L) (mean (SD))	2.52 (1.49)	2.16 (1.17)	0.7615 [0.6301, 0.9203]	0.00%	< 0.01
CRP(mg/L) (mean (SD))	24.27 (29.00)	36.62 (44.08)	1.0096 [1.0049, 1.0142]	0.23%	< 0.001
Calcium(mmol/L) (mean (SD))	2.36 (0.74)	2.26 (0.15)	0.0298 [0.0069, 0.1277]	15.62%	< 0.001
Sodium(mmol/L) (mean (SD))	137.36 (34.59)	135.82 (2.70)	0.9269[0.8658, 0.9923]	1.05%	< 0.05
CK(U/L) (mean (SD))	138.29 (188.98)	249.75 (796.52)	1.0007 [1.0002, 1.0011]	0.90%	< 0.01
LDH(U/L) (mean (SD))	344.18 (111.56)	416.28 (155.35)	1.0037 [1.0025, 1.0049]	1.65%	< 0.001
ALT(U/L) (mean (SD))	14.96 (22.12)	26.81 (38.56)	1.0115 [1.0047, 1.0184]	0.83%	< 0.001
IgM(g/L) (mean (SD))	1.38 (0.56)	1.60 (1.09)	1.5540 [1.2239, 1.9732]	4.35%	< 0.001
IgA(g/L) (mean (SD))	1.22 (0.73)	1.37 (0.57)	1.3044 [1.0283, 1.6545]	4.35%	< 0.05
Fer(ng/mL) (mean (SD))	114.36 (82.75)	210.50 (314.19)	1.0045 [1.0029, 1.0061]	6.76%	< 0.001
CD3(%) (mean (SD))	1987.92 (993.36)	1741.90 (894.19)	0.9997 [0.9994, 0.9999]	11.11%	< 0.01
NLR (mean (SD))	2.95 (2.33)	3.73 (2.33)	1.1153 [1.0470, 1.1881]	0.00%	< 0.001
MPVLR (mean (SD))	4.63 (3.10)	5.72 (3.35)	1.0746 [1.0289, 1.1223]	0.30%	< 0.01
PLR (mean (SD))	140.38 (79.19)	165.00 (80.14)	1.0028 [1.0010, 1.0047]	0.00%	< 0.01

Table 2. Significant Associations Identified in the Univariate Analysis of the Entire Dataset. *Abbreviations: refractory *Mycoplasma pneumoniae* pneumonia (RMPP); severe Mycoplasma pneumoniae pneumonia (SMPP); lymphocyte count (L); C-reactive protein (CRP); C-reactive protein (CRP); creatine kinase (CK); lactate dehydrogenase (LDH); alanine aminotransferase (ALT); immunoglobulin M (IgM); immunoglobulin A (IgA); ferritin (Fer); cluster of differentiation 3 (CD3);neutrophil-to-lymphocyte ratio (NLR); mean platelet volume to lymphocyte ratio (MPVLR); platelet-to-lymphocyte ratio (PLR);.

treatment before hospitalization, SMPP, LDH, NLR, ALT, peak fever, and extensive lung consolidation. Although the recall of the RF model was lower than that of XGBoost, other performance metrics were still satisfactory. Therefore, we included the SHAP summary plot for the RF model in Figure S1 for reference. Additionally, a permutation importance analysis is provided in Figure S2 to ensure the reliability of the results. The results showed that both XGBoost and RF consistently highlighted key variables—such as external macrolide treatment before hospitalization, SMPP, ALT, and extensive lung consolidation—across both SHAP importance and permutation importance analyses, reinforcing their significance in predicting RMPP.

To aid readers in understanding this complex machine learning model, we presented decision thresholds for these variables in Fig. 3. When the SHAP value of a specific feature exceeds zero, the decision tends towards the RMPP class, and vice versa for the non-RMPP class. For example, in Fig. 3A, if the actual value of NLR is < 3.51, the decision tends towards the non-RMPP class; if it > 3.51, the decision tends towards the RMPP class. Although this is a predictive model that cannot be directly interpreted, we have packaged it and uploaded it to GitHub (https://github.com/yuhan-coder/RMPP/tree/main), along with instructions for use, for readers to conveniently download and use.

In the summary plot, four variables showed a nonlinear impact on RMPP prediction: MPVLR, serum calcium, lymphocytes, and Fer. Their values corresponded to varying RMPP likelihoods—increasing or decreasing values correlated with RMPP likelihood, while intermediate values reduced it. However, serum calcium exhibited an inverse relationship; detailed information is available in Fig. 3.

Discussion

This study contributes to a deeper understanding of MPP, particularly by identifying risk factors associated with the progression to RMPP. The XGBoost algorithm was recognized as the most effective in predicting RMPP

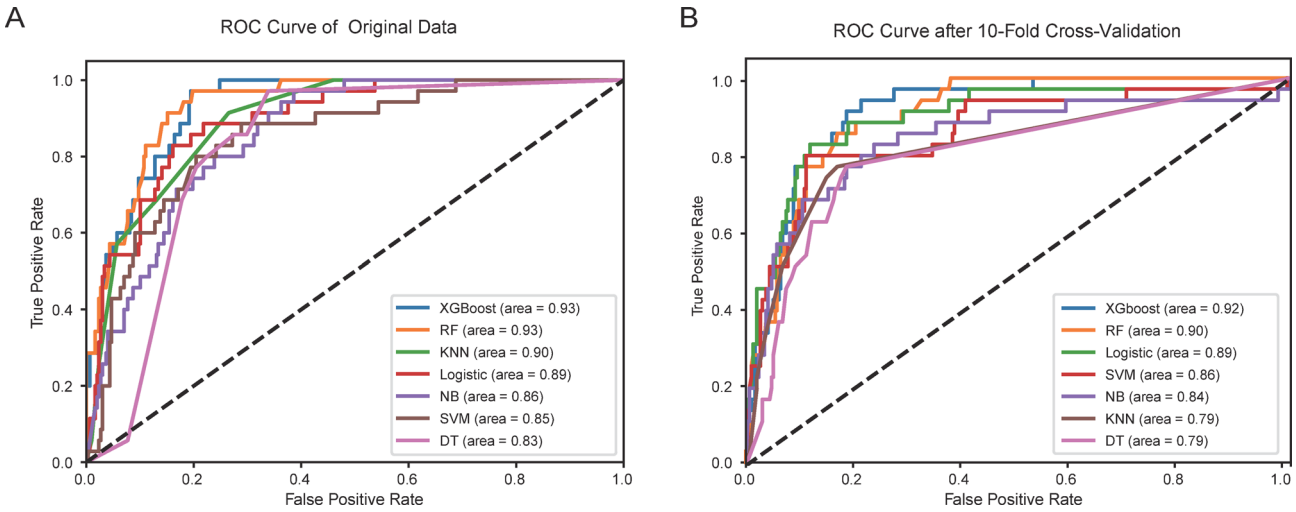

Fig. 1. **A:** Receiver Operating Characteristic (ROC) Curve Comparison of Machine Learning Models of the Original Data for RMPP Prediction. The graph presents the ROC curves of various machine learning models evaluated on the original testing dataset for RMPP prediction.The AUC values for each model are as follows: XGBoost attained the highest AUC of 0.93, closely followed by RF at 0.93. Logistic regression achieved an AUC of 0.89, KNN presented an AUC of 0.90, while SVM demonstrated an AUC of 0.85. NB and DT models performed at AUC values of 0.86 and 0.83, respectively. These AUC values reflect the predictive capacity of each model in diagnosing RMPP, with higher values indicating better predictive performance. **B**: Receiver Operating Characteristic (ROC) Curve Comparison of Machine Learning Models After tenfold Cross-Validation for RMPP Prediction. The graph illustrates the ROC Curve generated through tenfold cross-validation on the training dataset, comparing the predictive performance of various machine learning models for RMPP. XGBoost performed the best, achieving an AUC value of 0.92. Logistic regression and RF had AUC values of 0.89 and 0.90, respectively. KNN and SVM scored AUCs of 0.79 and 0.86, respectively. NB and DT models had AUCs of 0.84 and 0.79. Higher AUC values indicate better performance of the models in predicting RMPP.

	In the original data[#]					After tenfold cross-validation				
Model	AUC	Accuracy	Precision	Recall	F1Score	AUC	Accuracy	Precision	Recall	F1Score
XGboost	0.93	0.80	0.34	0.97	0.51	0.92	0.91	0.50	0.36	0.46
Logistic	0.89	0.85	0.39	0.74	0.51	0.89	0.91	0.68	0.45	0.41
KNN	0.90	0.85	0.38	0.69	0.49	0.79	0.90	0.55	0.10	0.25
RF	0.93	0.85	0.40	0.89	0.55	0.90	0.90	0.45	0.25	0.29
SVM	0.85	0.80	0.31	0.77	0.45	0.86	0.91	0.63	0.37	0.43
NB	0.86	0.76	0.28	0.77	0.41	0.84	0.88	0.47	0.66	0.48
DT	0.83	0.78	0.30	0.80	0.43	0.79	0.86	0.45	0.28	0.37

Table 3. Establishment and Evaluation of the Machine Learning Model. *Abbreviations: eXtreme Gradient Boosting (XGBoost); random forest (RF); naive bayes (NB); support vector machine (SVM); decision tree (DT) and K-nearest neighbors (KNN). [#]The original data were evaluated on the testing set, whereas tenfold cross-validation was performed on the training set.

by comparing seven machine learning models. The sensitivity analysis further validated the model's reliability. To improve our understanding of the complex machine learning model for RMPP, we used SHAP to interpret and simplify the XGBoost results, identifying eight key predictive factors and their decision thresholds, and highlighting their clinical priority for RMPP prediction. Finally, the prediction model was made available for download and utilization.

Our research has identified 29 predictive factors associated with RMPP. Many of these factors have been extensively documented in the literature, including CRP, LDH, ALT, age, duration of fever in days, and lymphocytes[30–32]. The alignment with established evidence supports the reliability of our findings. In addition, we uncovered some relevant composite variables that have received limited attention in previous studies, such as MPVLR, PLR, and NLR[33,34], which have gained recognition in recent years as sensitive markers for assessing inflammation. Subsequently, we used SHAP to interpret XGBoost results and identified 8 key variables, including during fever days and macrolide treatment before hospitalization, SMPP, LDH, NLR, ALT, peak fever, and extensive lung consolidation.

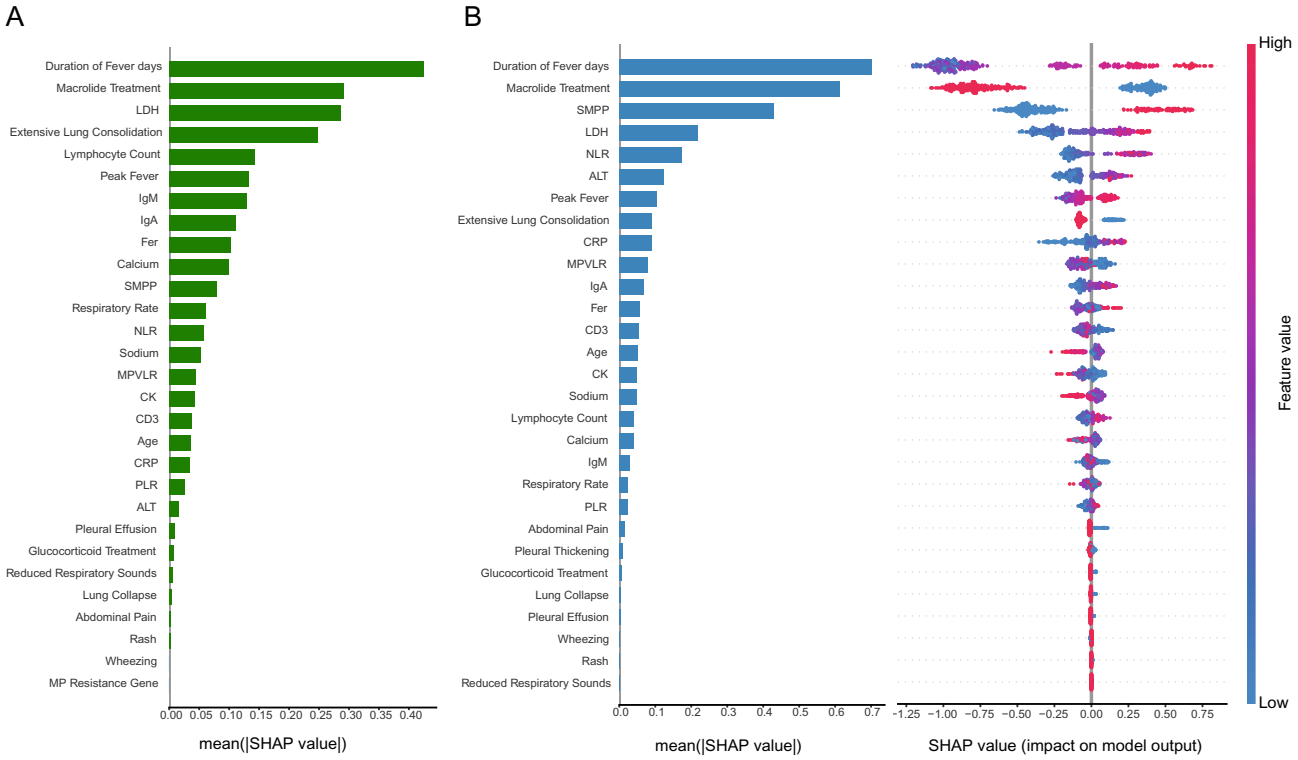

Fig. 2. **A:** SHAP Ranking Results for Predictive Model of RMPP in Sensitivity Analysis. The figure illustrates the SHAP ranking outcomes for the predictive model of RMPP during the sensitivity analysis of XGBoost Model 1. Features are arranged by their contribution to predicting RMPP, with the X-axis representing the SHAP values corresponding to each feature. Noteworthy features are positioned at the top of the chart, indicating their significant influence on predicting RMPP. **B**: SHAP Feature Importance Ranking for the XGBoost Model Predicting RMPP. The figure on the left displays the SHAP-based feature importance rankings for the XGBoost model based on the training dataset. Features are ranked based on their contribution to predicting RMPP, with the X-axis displaying the SHAP values for each feature. The right side shows the direction and magnitude of the effect, while the color gradient, from blue (low value) to red (high value), represents the original feature value distribution.

Consistent with a nomogram study predicting RMPP, fever duration and NLR were identified as key predictive factors[14], reflecting the body's inflammatory state. As fever duration and macrolide treatment before hospitalization were the two most significant indicators in our model, we recommend early assessment of macrolide response to enable the timely identification of RMPP. If children diagnosed with MPP show no response to macrolide treatment within the initial three days (e.g., persistent high fever or failure of inflammatory markers such as procalcitonin to decrease)[35], increased vigilance for RMPP progression is warranted. In addition, NLR reflects the balance between acute and chronic inflammation and adaptive immunity[36]. Reports suggest that an early increase after acute physiological stress could serve as an acute stress marker earlier than other laboratory parameters (e.g., white blood cell count, CRP). In addition to inflammation-related variables, this study also identified other factors with significant predictive value.

SMPP is a key predictor of RMPP. As a composite indicator, it reflects disease severity, indirectly pointing to tissue damage, heightened inflammatory responses, and immune dysregulation. This observation aligns closely with the underlying pathogenic mechanisms of RMPP[37]. In a previous machine learning study predicting the risk of hepatic injury in SMPP[38], ALT was identified as a key contributing factor. Interestingly, our study also recognized ALT as a significant predictor of RMPP. This consistent predictive value of ALT in both SMPP and RMPP underscores the significance of extrapulmonary injury markers in forecasting poor outcomes in MPP. Additionally, LDH held significant positions in our predictive model. LDH, found ubiquitously in cellular cytoplasm, typically indicates tissue damage when elevated, with higher levels correlating to increased RMPP severity[39,40]. Some studies even suggest LDH as a biomarker to initiate glucocorticoid therapy[41,42].

Unlike other machine learning studies on MPP, we not only focused on the overall predictive performance of the models but also emphasized the contribution of individual variables and their nonlinear trends. Many researchers argue that the emergence of macrolide-resistant strains in MP might escalate bacterial burden, and potentially contribute to the onset and progression of RMPP[37,43]. However, our sensitivity analysis, consistent with other studies, suggests that macrolide resistance may not have a significant correlation with disease severity[44,45]. Within our dataset, the incidence of MP resistance in MP stood at 90% and the rigorous statistical analysis revealed no statistically significant association between MP resistance and RMPP occurrence ($P = 0.88$). These findings were consistently supported by leave-out model assessments and variable importance rankings,

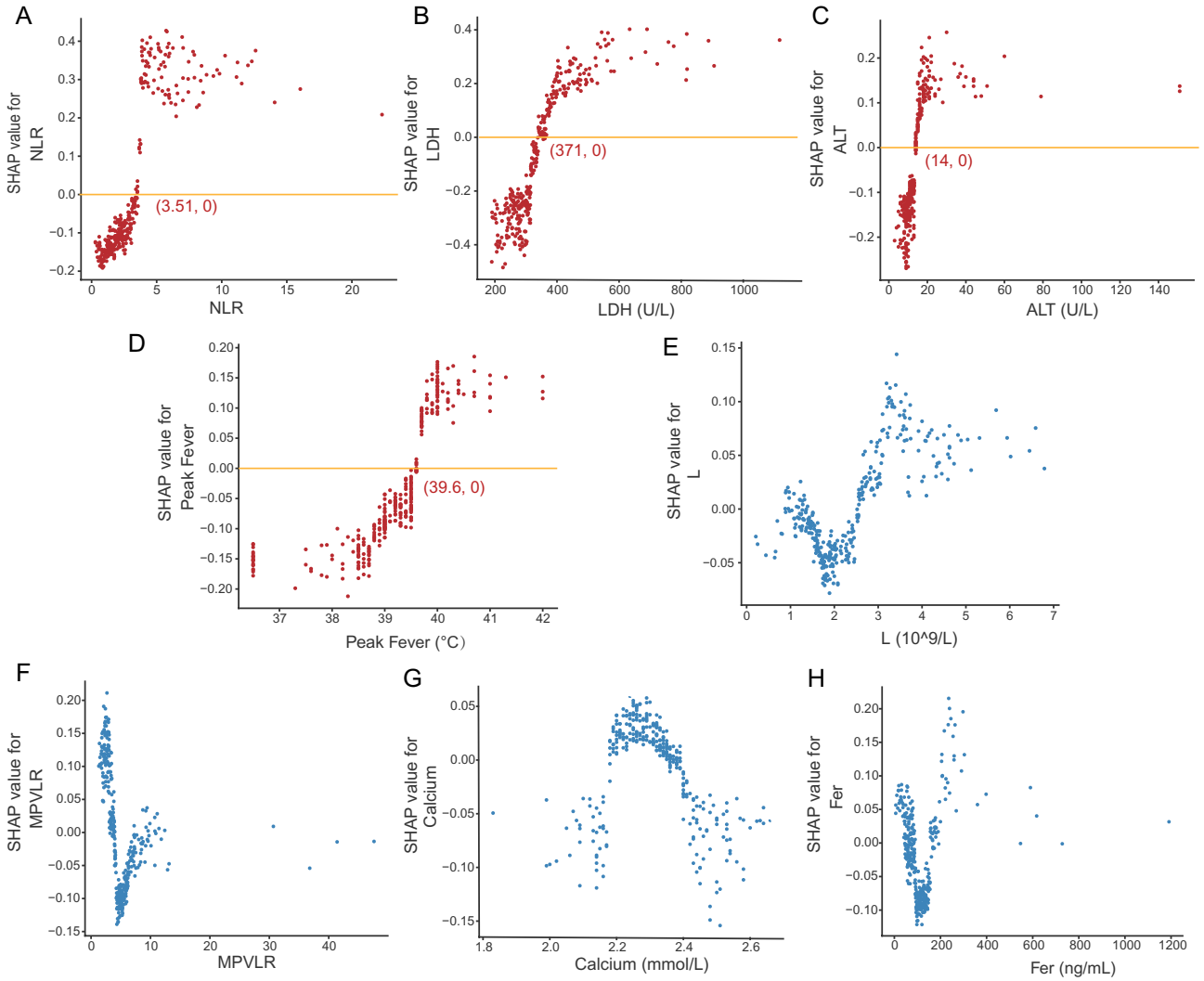

Fig. 3. SHAP Dependence Plot for Nonlinear Relationship Variable in XGBoost. The SHAP dependence plot visualizes the impact of a specific feature on the single variate (LDH/ NLR/ALT/peak fever/calcium/ lymphocytes/Fer/MPVLR) generated by the XGBoost model. Each dependency plot illustrates how an individual feature influences the predictive model's output, with each point representing a single individual. SHAP values are plotted on the y-axis, while actual values are on the x-axis. Annotated points indicate the critical value, where a specific feature's SHAP value exceeding zero steers the decision towards the EMPP class. For example, in **A**, if the actual value of A (NLR) is < 3.51, the decision tends towards the non-RMPP class; if it > 3.51, the decision tends towards the RMPP class.

emphasizing the limited predictive capacity of MP resistance in forecasting RMPP incidence. These pieces of evidence challenge the previously assumed strong link between macrolide resistance and the severity of MP infections. In this highly resistant context, the predictive value of macrolide resistance in determining RMPP development might be substantially diminished. Future studies could further investigate the variations and evolutionary trends of MP during its periodic outbreaks to explore its molecular mechanisms associated with RMPP.

Finally, this study has discovered several indicators that show a non-linear correlation with RMPP, which may be related to the pathogenesis of RMPP. Abnormalities in MPVLR, lymphocyte ratio, and Fer levels (whether increased or decreased) are associated with RMPP. Typically, neutrophils play a crucial role as the first line of defense against infection in the immune response to MP. After MP infection, interactions among neutrophils, lymphocytes, platelets, and others mediate the relationship between inflammation and thrombosis. Excessive inflammation may induce lymphocyte apoptosis, leading to a decrease in their numbers[36]. Therefore, we need to not only focus on the transient elevation of immune cells during the immune response but also be attentive to the depletion status of various immune cells and factors after the immune storm[37]. Furthermore, serum ferritin is an inflammatory marker, and its levels are typically expected to rise during acute inflammatory conditions. The relationship mentioned may need further validation and clarification after controlling for confounding factors.

Strengths and limitations

Recent studies in China have applied machine learning to the early detection and risk prediction of MPP, including models based on organ damage[38] related to SMPP and routine blood tests[20], all demonstrating high accuracy. However, this study offers several key advantages. First, more comprehensive clinical data were integrated, including multidimensional information such as demographic characteristics, symptoms, laboratory results, and imaging findings, leading to a more representative predictive model. Second, various machine learning algorithms were systematically compared, and sensitivity analysis, along with summary plots, highlighted nonlinear patterns in several key variables related to RMPP risk. Third, SHAP analysis highlighted key clinically interpretable features, clarified decision thresholds, and ultimately developed a simplified, high-performance model with eight essential variables. Finally, the final model was made publicly available, enabling clinicians to use it in practice and providing researchers with further model integration and secondary data mining.

Nevertheless, this study faces certain limitations. Firstly, despite employing multiple imputations, handling missing data might not have been entirely effective, potentially resulting in information loss or bias. Secondly, sample imbalance presents a significant challenge, possibly inducing model bias during training and limiting performance in specific categories. These challenges might impact model stability and generalization capability, necessitating comprehensive consideration and resolution in future studies[46]. Lastly, this study is a single-center retrospective analysis. Although internal validation was performed to ensure the reliability of the results, the lack of external validation may limit its generalizability to other populations. Further validation and extension are necessary through larger multi-center data verification and extrapolation in future research endeavors.

Conclusion

Through the application of machine learning, key factors associated with the development and progression of RMPP, as well as non-linear relationship factors potentially linked to its onset, have been successfully identified. These findings offer valuable insights for future clinical practices, potentially enhancing preventive and treatment approaches for RMPP.

Data availability

The model, key variable data, and usage instructions for the article can be obtained from GitHub (https://github.com/yuhan-coder/RMPP/tree/main). For comprehensive individual clinical data, due to confidentiality restrictions, requests should be directed to the corresponding author.

Received: 29 July 2024; Accepted: 16 May 2025
Published online: 23 May 2025

References

1. Yan, C., Sun, H. & Zhao, H. Latest surveillance data on mycoplasma pneumoniae infections in children, suggesting a new epidemic occurring in Beijing. *J Clin Microbiol.* **54**(5), 1400–1401 (2016).
2. Lind, K., Benzon, M. W., Jensen, J. S. & Clyde, W. A. A seroepidemiological study of Mycoplasma pneumoniae infections in Denmark over the 50-year period 1946–1995. *Eur. J. Epidemiol.* **13**(5), 581–586 (1997).
3. Tong, L., Huang, S., Zheng, C., Zhang, Y. & Chen, Z. Refractory mycoplasma pneumoniae pneumonia in children: Early recognition and management. *J Clin Med.* **11**(10), 2824 (2022).
4. Biagi, C. et al. Pulmonary and extrapulmonary manifestations in hospitalized children with mycoplasma pneumoniae infection. *Microorganisms.* **9**(12), 2553 (2021).
5. Inamura, N. et al. Management of refractory Mycoplasma pneumoniae pneumonia: Utility of measuring serum lactate dehydrogenase level. *J. Infect. Chemother.* **20**(4), 270–273 (2014).
6. China National Health Commission (2023). Guidelines for the diagnosis and treatment of Mycoplasma pneumoniae pneumonia in children (2023).) [EB/OL]. [2023–02–15]. Available at: Available at: http://www.nhc.gov.cn/yzygj/s7659/202302/8536e7db5cc7443eba13601e58d58861.shtml.
7. Ding, Y. et al. High expression of HMGB1 in children with refractory Mycoplasma pneumoniae pneumonia. *BMC Infect Dis.* **29**(18), 439 (2018).
8. Gao, L. W. et al. The epidemiology of paediatric Mycoplasma pneumoniae pneumonia in North China: 2006 to 2016. *Epidemiol Infect.* **147**, e192 (2019).
9. Tsai, T. A., Tsai, C. K., Kuo, K. C. & Yu, H. R. Rational stepwise approach for Mycoplasma pneumoniae pneumonia in children. *J. Microbiol. Immunol. Infect.* **54**(4), 557–565 (2021).
10. Ling, Y., Ning, J. & Xu, Y. Explore the predictive value of peripheral blood cell parameters in refractory mycoplasma pneumoniae pneumonia in children over 6 years old. *Front Pediatr.* **12**(9), 659677 (2021).
11. Li, G. et al. High co-expression of TNF-α and CARDS toxin is a good predictor for refractory Mycoplasma pneumoniae pneumonia. *Mol Med.* **9**(25), 38 (2019).
12. Kim, S. H., Lee, E., Song, E. S. & Lee, Y. Y. Clinical Significance of Pleural Effusion in Mycoplasma pneumoniae Pneumonia in Children. *Pathogens* **10**(9), 1075. https://doi.org/10.3390/pathogens10091075 (2021).
13. Wang, X. et al. Necrotizing pneumonia caused by refractory Mycoplasma pneumonia pneumonia in children. *World J. Pediatr.* **14**(4), 344–349 (2018).
14. Pei, H. & Luo, H. Predictive clinical indicators of refractory Mycoplasma pneumoniae pneumonia in children: A retrospective cohort study. *Med. (Baltimore)* **103**(34), e39375 (2024).
15. Lin, J. D. et al. Comparison between machine learning and multiple linear regression to identify abnormal thallium myocardial perfusion scan in Chinese Type 2 diabetes. *Diagnostics (Basel).* **12**(7), 1619 (2022).
16. Huang, L. Y. et al. Comparing multiple linear regression and machine learning in predicting diabetic urine albumin-creatinine ratio in a 4-year follow-up study. *J. Clin. Med.* **11**(13), 3661 (2022).
17. Liu, Y., Chen, P. H. C., Krause, J. & Peng, L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*. **322**(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489 (2019).
18. Zhuo, L. Y. et al. Predicting the severity of mycoplasma pneumoniae pneumonia in pediatric and adult patients: a multicenter study. *Sci Rep.* **14**(1), 22978 (2024).
19. Qian, Y. et al. Model based on the automated AI-driven CT quantification is effective for the diagnosis of refractory Mycoplasma pneumoniae pneumonia. *Sci. Rep.* **14**(1), 16172 (2024).

20. Peng, X. et al. A preliminary prediction model of pediatric Mycoplasma pneumoniae pneumonia based on routine blood parameters by using machine learning method. *BMC Infect. Dis.* **24**(1), 707 (2024).
21. Lundberg, S. & Lee, S. I. A unified approach to interpreting model predictions | Proceedings of the 31st International Conference on Neural Information Processing Systems. Guide Proceedings. https://doi.org/10.5555/3295222.3295230.
22. Ou, S. M. et al. Artificial Intelligence for risk prediction of rehospitalization with acute kidney injury in sepsis survivors. *J Pers. Med.* **12**(1), 43 (2022).
23. Ying, Y. H., Lee, W. L., Chi, Y. C., Chen, M. J. & Chang, K. Demographics, socioeconomic context, and the spread of infectious disease: The case of COVID-19. *Int. J. Environ. Res. Public Health.* **19**(4), 2206 (2022).
24. Rajput, D., Wang, W. J. & Chen, C. C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinform.* **14**(24), 48 (2023).
25. McIntosh, K. Community-acquired pneumonia in children. *New Engl. J. Med.* **346**, 429–437. https://doi.org/10.1056/NEJMra011994 (2002).
26. Zhong, H. et al. Analysis of clinical characteristics and risk factors of plastic bronchitis in children with mycoplasma pneumoniae pneumonia. *Front. Pediatr.* **18**(9), 735093 (2021).
27. Yang, L. et al. Clinical features and risk factors of plastic bronchitis caused by Mycoplasma pneumoniae pneumonia in children. *BMC Pulm. Med.* **23**(23), 468 (2023).
28. Marzi, C. et al. Collinearity and dimensionality reduction in radiomics: Effect of preprocessing parameters in hypertrophic cardiomyopathy magnetic resonance T1 and T2 mapping. *Bioengineering (Basel).* **10**(1), 80 (2023).
29. Qu, W. et al. Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging. *Int. J. Comput. Assist. Radiol. Surg.* **15**(12), 2041–2048 (2020).
30. Li, M. et al. Recognition of refractory Mycoplasma pneumoniae pneumonia among Myocoplasma pneumoniae pneumonia in hospitalized children: Development and validation of a predictive nomogram model. *BMC Pulm. Med.* **10**(23), 383 (2023).
31. Zhao, L. et al. Development and validation of a nomogram to predict plastic bronchitis in children with refractory Mycoplasma pneumoniae pneumonia. *BMC Pulm. Med.* **27**(22), 253 (2022).
32. Zhu, J. et al. Predictive value of chemokines (CCL 2) in bronchoalveolar lavage fluid for refractory mycoplasma pneumonia in children. *Ital. J. Pediatr.* **23**(49), 125 (2023).
33. Ling, Y., Ning, J. & Xu, Y. Explore the Predictive Value of Peripheral Blood Cell Parameters in Refractory Mycoplasma pneumoniae Pneumonia in Children Over 6 Years Old. *Front. Pediat.* **9**, 659677. https://doi.org/10.3389/fped.2021.659677 (2021).
34. Liu, G. et al. A lung ultrasound-based nomogram for the prediction of refractory mycoplasma pneumoniae pneumonia in hospitalized children. *Infect. Drug. Resist.* **31**(15), 6343–6355 (2022).
35. Chang, C. H. et al. Epidemiology and clinical manifestations of children with macrolide-resistant Mycoplasma pneumoniae pneumonia in Southern Taiwan. *Pediatr. Neonatol.* **62**(5), 536–542 (2021).
36. Li, D. et al. Neutrophil-to-lymphocyte ratio as a predictor of poor outcomes of Mycoplasma pneumoniae pneumonia. *Front. Immunol.* **14**, 1302702 (2023).
37. Tong, L., Huang, S., Zheng, C., Zhang, Y. & Chen, Z. Refractory Mycoplasma pneumoniae Pneumonia in Children: Early Recognition and Management. *J. Clin. Med.* **11**(10), 2824. https://doi.org/10.3390/jcm11102824 (2022).
38. He, B. et al. Development of machine learning-based differential diagnosis model and risk prediction model of organ damage for severe Mycoplasma pneumoniae pneumonia in children. *Sci Rep.* **15**(1), 9431 (2025).
39. Zheng, X. X., Lin, J. L. & Dai, J. H. Value of lactate dehydrogenase in predicting refractory Mycoplasma pneumoniae pneumonia in children: an evaluation based on decision curve analysis and dose-response analysis. *Zhongguo Dang Dai Er Ke Za Zhi* **22**(2), 112–117 (2020).
40. Wang, S. et al. Diagnostic value of serum LDH in children with refractory Mycoplasma pneumoniae pneumoniae: A systematic review and meta-analysis. *Front Pediatr.* **11**, 1094118 (2023).
41. Liu, T. Y. et al. Serum lactate dehydrogenase isoenzymes 4 plus 5 is a better biomarker than total lactate dehydrogenase for refractory Mycoplasma pneumoniae pneumonia in children. *Pediatr Neonatol.* **59**(5), 501–506 (2018).
42. Xu, J. J. & Shu, L. H. Clinical characteristics of refractory Mycoplasma pneumoniae pneumonia in children. *Zhongguo Dang Dai Er Ke Za Zhi* **20**(1), 37–42 (2018).
43. Chen, Y., Li, L., Wang, C., Zhang, Y. & Zhou, Y. Necrotizing Pneumonia in children: Early recognition and management. *J Clin Med.* **12**(6), 2256 (2023).
44. Yang, T. I. et al. Mycoplasma pneumoniae in pediatric patients: Do macrolide-resistance and/or delayed treatment matter?. *J. Microbiol. Immunol. Infect.* **52**(2), 329–335 (2019).
45. Kutty, P. K. et al. Mycoplasma pneumoniae among children hospitalized with community-acquired pneumonia. *Clin. Infect. Dis.* **68**(1), 5–12 (2019).
46. Abbas, Q., Malik, K. M., Saudagar, A. K. J. & Khan, M. B. Context-aggregator: An approach of loss- and class imbalance-aware aggregation in federated learning. *Comput. Biol. Med.* **163**, 107167 (2023).

Acknowledgements

We thank all participants and staff of this study and the physicians at the Tianjin Children's Hospital affiliated with Tianjin University.

Author contributions

YJ. LL, XW., YW., XW., and YZ. designed the study and drafted the manuscript. YJ. prepared all figures. All authors reviewed the manuscript. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Funding

This study was financially supported by the Tianjin Municipal Health Commission Key Discipline Special Fund (TJWJ2022XK038) and Tianjin Key Medical Discipline (Specialty) Construction Project (TJYXZDXK-040A).

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

This study was conducted in accordance with the principles of the Helsinki Declaration. It was approved by

the Ethics Committee of Tianjin Children's Hospital (No. 022-LXKY-004). All procedures involving human participants were in line with the ethical standards of the institutional and/or national research committee, and written informed consent was obtained from individual participants and their legal guardians.

Consent for publication

Not applicable.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-02962-4.

Correspondence and requests for materials should be addressed to Y.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.