BMC Infectious Diseases

RESEARCH

Open Access

Machine learning-based risk prediction model for pertussis in children: a multicenter retrospective study

Juan Xie[1†], Run-wei Ma[2†], Yu-jing Feng[3], Yuan Qiao[4], Hong-yan Zhu[5], Xing-ping Tao[6], Wen-juan Chen[7], Cong-yun Liu[8], Tan Li[9], Kai Liu[10*] and Li-ming Cheng[1*]

Abstract

Background Pertussis is a highly contagious respiratory disease. Even though vaccination has reduced the incidence, cases have resurfaced in certain regions due to immune escape and waning vaccine efficacy. Identifying high-risk patients to mitigate transmission and avert complications promptly is imperative. Nevertheless, the current diagnostic methods, including PCR and bacterial culture, are time-consuming and expensive. Some studies have attempted to develop risk prediction models based on multivariate data, but their performance can be improved. Therefore, this study aims to further optimize and expand the risk assessment tool to more efficiently identify high-risk individuals and compensate for the shortcomings of existing diagnostic methods.

Objective The aim of this study was to develop a pertussis risk prediction model that is both efficient and has good generalization ability, applicable to different datasets. The model was constructed using machine learning techniques based on multicenter data and screened for key features. The performance and generalization ability of the model were evaluated by deploying it on an online platform. At the same time, this study aims to provide a rapid and accurate auxiliary diagnostic tool for clinical practice to help identify high-risk patients in a timely manner, optimize early intervention strategies, reduce the risk of complications and reduce transmission, thereby improving the efficiency of public health management.

Methods First, data from 1085 suspected pertussis patients from 7 centers were collected, and ten key features were analyzed using the lasso regression and Boruta algorithm: PDW-MPV-RATIO, SII, white blood cells, platelet distribution width, mean platelet volume, lymphocytes, cough duration, vaccination, fever, and lytic lymphocytes.Eight models were then trained and validated to assess their performance and to confirm their generalization ability with external datasets based on these features. Finally, an online platform was constructed for clinicians to use the models in real time.

Results The random forest model demonstrated excellent discrimination ability in the validation set, with an AUC of 0.98, and an AUC of 0.97 in the external validation set. Calibration curve and decision curve analysis showed that the model had high accuracy in predicting low-to-medium risk patients, which could help clinicians avoid

[†]Juan Xie and Run-wei Ma contributed equally as co-first authors.

*Correspondence:
Kai Liu
ynkmlk@foxmail.com
Li-ming Cheng
Medcheng@126.com
Full list of author information is available at the end of the article

unnecessary interventions, especially in resource-limited settings. The application of this model can help optimize the early identification and management of high-risk patients and improve clinical decision-making.

Conclusion The pertussis prediction model devised in this study was validated using multicenter data, exhibited high prediction performance, and was successfully implemented online. Future research should broaden the data sources and incorporate dynamic data to enhance the model's accuracy and applicability.

Keywords Public health, PDW-MPV-RATIO, SII, Calibration curves, Lasso regression, Random forest, Online deployment, Pertussis

Introduction

Pertussis, caused by Bordetella pertussis, is a highly contagious respiratory disease and predominantly affects children, particularly infants and young children who have not received their vaccinations. Although global vaccination has significantly reduced the incidence of pertussis, the resurgence of pertussis in some areas in recent years is of great concern. The resurgence of pertussis may be linked to various factors, including the pathogen's genetic variation, the progressive decline in vaccine efficacy, and insufficient vaccination coverage [1–3]. The timely and accurate diagnosis of pertussis is of the utmost importance, particularly in infants and young children, as the symptoms of the disease initially resemble those of typical respiratory infections. A delay in diagnosis may result in severe complications or even mortality [4, 5].

The primary diagnostic methods for pertussis are clinical manifestations and laboratory tests, such as bacterial culture, PCR, and serologic assays. Despite their significant role in diagnosis confirmation, these methods have clear limitations: bacterial culture is time-consuming and has low sensitivity. At the same time, PCR testing is expensive and requires sophisticated technical facilities, which restricts its use in specific clinical circumstances [6, 7]. Consequently, creating an early risk prediction tool based on clinical and laboratory data is essential. This tool will enable the earlier identification of at-risk patients, thereby enhancing the timeliness of clinical interventions and reducing misdiagnosis [8, 9].

The application of machine learning techniques in the medical field has progressively expanded in recent years, particularly in predicting disease risk, demonstrating significant potential [10–12]. Machine learning models can capture the nonlinear relationship between variables and provide more precise disease risk prediction by integrating and analyzing complex multidimensional data [13, 14]. Machine learning models have been demonstrated to substantially enhance the early identification of respiratory diseases in the context of risk prediction [15, 16]. At present, although progress has been made in pertussis prediction research, there are still limitations. The machine learning algorithm developed by Daluwatte et al. can effectively identify undiagnosed or misdiagnosed cases in electronic health records, but its recognition rate and accuracy are affected by the quality and diversity of the data set, limiting its generalizability in different populations and healthcare systems [17]. Similarly, Tozzi et al.'s data-driven algorithm improves the accuracy of early diagnosis of whooping cough, but the algorithm depends on the availability and completeness of the data and may not perform well in different environments [18]. Therefore, this field needs multi-center data and more complex feature selection models to improve the universality and robustness of the prediction. [19, 20]

This study addresses the current deficiency in pertussis risk prediction by developing and validating a machine-learning model based on multicenter data. This model will offer clinicians a precise and efficient instrument for the early detection of pertussis. This tool not only enhances clinical decision-making and minimizes misdiagnosis and omission, but it also establishes a scientific foundation for public health prevention and control strategies, particularly in managing the pertussis recurrence trend [21, 22].

Methods

Data sources and study design

Data for the study came from seven medical centers in Yunnan Province: Kunming Children's Hospital, Wenshan Maternal and Child Health Hospital, Chuxiong People's Hospital, Qujing Maternal and Child Health Hospital, Yuxi Children Hospital, Kaiyuan Children Hospital and Baoshan People's Hospital.

Between October 2023 and September 2024, we collected data on 1085 hospitalized patients with suspected whooping cough from 7 hospitals. This study encompassed all patients with suspected pertussis, who were subsequently validated or excluded by clinical presentation and laboratory testing.An Electronic Medical Record (EMR) is a digital tool used by hospitals to manage patient medical information. It converts traditional paper medical records into electronic form, recording information such as the patient's medical history, diagnosis, treatment, medication and test results. The same Electronic Medical Record (EMR) is

used in our seven centers. By using the same platform, the centers can more efficiently share patients' medical information, achieve cross-center collaboration and data analysis, improve the overall quality of medical services and the accuracy of decision-making. The hospital's electronic medical record (EMR) system is used to extract data on all patients in the study, including comprehensive clinical and laboratory examination information. The study was approved by the Ethics Committee, and all patients signed informed consent forms.

Control of bias

Several measures were implemented in this investigation to mitigate bias in multicenter studies. First, all data were extracted from the electronic medical record systems of each hospital through a standardized process to ensure data consistency. Specifically, each center used the same format and methods for data collection to ensure the uniformity and integrity of patient information. In addition, all data underwent a consistent quality control process to ensure data accuracy and consistency. These measures effectively reduced the bias that may be caused by different data collection methods. Secondly, the sample size of each center was balanced to prevent the model from being overly influenced by any one center.

Study population

Criteria for inclusion

1. Patients under the age of 18 who are admitted;

2. Possess comprehensive clinical and laboratory examination data;

3. Patients who satisfy the diagnostic criteria for pertussis are diagnosed by the requirements established by the National Health and Wellness Commission of China. Typical symptoms and the disease duration are included in the clinical criteria. Typical symptoms include frequent coughing, rales, vomiting, and, in some cases, cyanosis or apnea, particularly in neonates. The disease is divided into three phases: the kata phase (characterized by mild cough, fever, and other symptoms), the erratic cough phase (characterized by severe cough, rales, and vomiting), and the recovery phase (characterized by a progressive reduction in cough). Blood tests and pathogenetic testing comprise laboratory criteria. PCR testing, serologic antibody testing, or nasopharyngeal probe bacterial culture comprise pathogenetic testing. As blood tests indicate, an elevated lymphocyte ratio is typically observed in conjunction with leukocytosis, typically over 15,000–30,000/mm^3.

Criteria for exclusion

1. Patients who present with paroxysmal cough, dyspnea, or other symptoms similar to those of whooping cough. Examples include other Bacterial pneumonia, Mycoplasma infections, and viral respiratory infections (e.g.,adenovirus and respiratory syncytial virus infections).Bronchial asthma and allergic wheezing.
2. Patients with incomplete data: Since incomplete data may affect the accuracy of model construction, all patients with incomplete data in this study were excluded. This processing is intended to ensure the integrity and reliability of the data during model training and analysis.

Variable definition and data acquisition

The electronic medical record system of the hospital is utilized to extract clinical data for all patients. The variables that were collected are as follows:

Demographic information, including age, sex, height, weight, BMI, place of residence (urban/rural), and pregnancy status;

Clinical symptoms include paroxysmal cough, rales, regurgitation, cyanosis, apnea, fever, dyspnea, heart rate, temperature, and a prolonged cough.

Laboratory indices include the following: white blood cell count (WBC), lymphocytosis, neutrophil count, platelet count, lymphocyte count, monocyte count, erythrocyte hematocrit, lymphocyte percentage, platelet distribution width (PDW), mean platelet volume (MPV), serum albumin (Serum Albumin), C-reactive protein (CRP), PDW to MPV ratio (PDW-MPV-Ratio), neutrophil to lymphocyte ratio (NLR), systemic inflammation index (SII), lymphocyte to monocyte ratio (LMR), platelet-to-lymphocyte ratio (PLR), nutritional index (PNI), inflammation score (AISI), and inflammation-to-trophic ratio (CAR); and cleaved lymphocytes.

The Systemic Inflammation Index (SII) formula is as follows: SII = platelet count × neutrophil count/lymphocyte count, where platelet count, lymphocyte count, and neutrophil count are the three metrics of peripheral blood. These parameters were determined using a fully automated hematology analyzer on blood samples.

The proportion of lymphocytes with cleaved or lobulated nuclei was determined by counting the cleaved lymphocytes (lymphocytes) identified through a peripheral blood smear and Wright-Giemsa staining. The lymphocytes were then observed under a microscope by an experienced pathologist.

Imaging and vaccination data: history of vaccination, radiographic findings, and contact with pertussis patients;

The severity of illness and treatment includes SpO2 level, antibiotic treatment, and disease severity at admission.

Standardized procedures were implemented to quantify all laboratory data. Categorical variables are represented as frequency and percentage, while continuous variables are defined as mean ± standard deviation.

Feature selection

The following methodologies were employed to identify critical predictive features that are associated with the risk of contracting pertussis:

1. Lasso regression: covariance is handled and the model is simplified by L1 regularization, and the optimal λ value is determined by tenfold cross-validation.
2. Boruta algorithm: a feature selection method that utilizes random forests to ascertain the relative significance of each feature. The intersection features Lasso and Boruta screened are ultimately employed in the subsequent model construction.

Model development and assessment

The following eight machine learning models were employed in this study to predict the risk of pertussis: logistic regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBM), XGBoost algorithm, Multi-Layer Perceptron (MLP), Naive Bayes, and Decision Tree. The model is trained using the training set, and the model's generalization ability is assessed using the validation set. The dataset was randomly divided into training and validation sets at an 8:2 ratio.

This study presents the results of the evaluation of the validation set, including the AUC (area under the subject operating characteristic curve), sensitivity, specificity, and other relevant performance indicators, to highlight the actual performance of the model in a simplified manner. These evaluation metrics comprehensively reflect the prediction ability and generalization performance of the model. In order to simplify the presentation of results in the main text, only the key evaluation results of the validation set are reported, and the data from the training set and external validation are listed in the appendix.

Assessment of the model

Evaluation of classification performance: AUC and ROC curves assess the model's classification performance.

Calibration curve: evaluates the degree to which the model's anticipated probabilities correspond to the actual outcomes;

Decision curve analysis (DCA) is employed to evaluate the clinical advantages of the model.

Grid search and tenfold cross-validation are employed to optimize the model's hyperparameters.

External validation

External validation was conducted to assess the model's generalizability using data from 85 patients from Kaiyuan Children's Hospital. The external dataset comprised 41 confirmed pertussis patients and 44 non-pertussis patients. Due to the limited sample size, SMOTE (Synthetic Minority Over-sampling Technique) was applied to augment the dataset. By generating synthetic samples of the minority class, SMOTE effectively balances the ratio of diseased and non-diseased people in the dataset while increasing the overall sample size. This method not only alleviates the problem of data imbalance, but also provides more training data for the model, which helps the model learn features more fully, thereby improving the accuracy and generalization ability of the prediction.

Model interpretation and online deployment

The SHAP (Shapley Additive Explanations) analysis was employed in this study to enhance the model's interpretability by revealing each feature's contribution to the model's prediction results. The SHAP scatterplot and dependency plot illustrate the significance of each feature. Ultimately, the study's findings were implemented via an online platform that enables physicians to acquire real-time pertussis risk prediction results by inputting the key characteristics of their patients.

Statistical analysis

R software (version 4.2.0) was employed to conduct all statistical analyses. The "glmnet" and "Boruta" packages were used to perform feature selection, the "caret" package was used to construct and evaluate the model, and Python was used to conduct SHAP analyses. The "DCA" package was used to conduct the decision curve analysis, while the "rms" package was utilized to generate calibration curves. The statistical significance of AUC differences among models was assessed using DeLong's test. Statistical significance was defined as $p < 0.05$, and all tests were two-sided.

Results

The flowchart illustrates the selection and data segmentation of suspected pertussis cases employed for developing and validating the model. After excluding incomplete data, 1,032 of the 1,085 suspected cases from

seven hospitals were eligible. Of these, 947 cases from six hospitals were utilized for model development, while 85 cases from one hospital were used for external validation. 652 of the 947 cases were identified as pertussis, while 295 were classified as nonpertussis. (S1).

Baseline characteristics of the survey population

The baseline characteristics of the study population are summarized in Table 1. In the research population, the mean age was 6.53 ± 2.95 years, with most participants being children. The mean height was 1.09 ± 0.26 m, the mean weight was 20.10 ± 10.75 kg, and the mean BMI was 16.11 ± 3.98. The mean value of PDWMPVRATIO was 1.21 ± 0.38, the mean value of PDW was 11.60% ± 3.00%, and the mean value of MPV was 9.59 ± 1.15 fL among the blood and inflammation indices. The mean SII was 1,065.71 ± 3,143.27, indicating substantial variability in individual inflammatory responses. The mean value of NLR was 1.67 ± 2.81, reflecting the diversity of inflammatory responses.

Regarding vaccination, 44.14% of the research subjects received four doses of the pertussis vaccine, which may impact the spread of the disease. The study subjects' mean duration of cough was 10.61 ± 7.39 days and the incidence of fever was 74.55%. The incidence of pertussis was not significantly influenced by gender, and the gender distribution was nearly equalized, with 55.12% males and 44.88% females.

Selection of features

This study employed Lasso regression and Boruta feature selection algorithms to identify critical predictive features for pertussis risk. Lasso regression reduces model complexity by regularizing the parameter λ. Characteristics such as PDW-MPV-RATIO and SII maintain high coefficients at large λ, indicating their significance, while the coefficients of certain features taper off or go to zero (Fig. 1A). Lasso's deviation plot (Fig. 1B) showed that the binomial deviation of the model decreased significantly as λ increased, indicating that the model fitting effect was improved. Boruta's algorithm further confirmed the importance of characteristics such as PDW-MPV-RATIO, SII, and WBC by comparing the random forest with shadow features, which demonstrated a strong predictive ability in pertussis risk prediction (Fig. 1C).

We chose variables with importance values of 10 or more for the results, enhancing the model's simplicity and usefulness, as Boruta's algorithm integrates many variables.The Euler diagram (Fig. 2A) demonstrates that the Lasso and Boruta algorithms prioritize features such as PDW-MPV-RATIO and SII. The regression coefficients (Odds Ratio, OR) and their 95% confidence intervals for each feature were illustrated in logistic regression

Table 1 Baseline characteristics of the study population

Characteristics	n (%), mean ± SD (range), or med [IQR]
Age (year)	6.53 ± 2.95
Height (m)	1.09 ± 0.26
Weight (kg)	20.10 ± 10.75
BMI	16.11 ± 3.98
PDWMPVRATIO	1.21 ± 0.38
SII	1065.71 ± 3143.27
NLR	1.67 ± 2.81
LMR	8.06 ± 7.59
PLR	99.12 ± 58.38
PNI	74.71 ± 143.82
AISI	367.98 ± 610.70
CAR	0.15 ± 0.29
WBC 10^9/L	11.48 ± 5.75
Neutrophil10^9/L	4.68 ± 4.79
Platelet10^9/L	357.92 ± 119.74
Lymphocyte10^9/L	8.47 ± 5.01
Monocyte10^9/L	0.71 ± 0.60
Hematocrit %	38.50 ± 9.18
Lymphocyte Percentage %	45.57 ± 18.15
Platelet Distribution Width %	11.60 ± 3.00
Mean Platelet Volume 10^−15L	9.59 ± 1.15
Serum Albumin g/dL	48.63 ± 142.66
CRP	4.15 ± 5.71
Season	6.71 ± 2.11
Cough days	10.61 ± 7.39
Heart rate bpm	114.58 ± 18.07
Temperature ℃	36.98 ± 9.82
SpO2	94.39 ± 2.70
Gender	Male 522 (55.12%)
	Female 425 (44.88%)
Vaccination dose	1: 57 (6.02%)
	2: 179 (18.9%),
	3: 73 (7.71%),
	4: 418 (44.14%),
	5: 220 (23.23%)
Cleaved lymphoctes	Yes 599 (63.25%)
	No 348 (36.75%)
Fever	Yes 706 (74.55%)
	No 241 (25.45%)
Paroxysmal cough	Yes 946 (99.89%)
	No 1 (0.11%)
Whooping cough	Yes 437 (46.15%)
	No 510 (53.85%)
Vomiting	Yes 404 (42.66%)
	No 543 (57.34%)
Cyanosis	Yes 22 (2.32%)
	No 925 (97.68%)

Table 1 (continued)

Characteristics	n (%), mean ± SD (range), or med [IQR]
Apnea	Yes 3 (0.32%)
	No 944 (99.68%),
Dyspnea	Yes 25 (2.85%),
	No 920 (97.15%),
Pertusiss	Yes 652 (68.85%)
	No 295 (31.15%),
X-ray	Yes 655 (69.17%)
	No 292 (30.83%),
Contact with Pertussis Patient	Yes 695 (73.39%)
	No 252 (26.61%)
AntibioticTreatment	Yes 931 (98.31%)
	No 16 (1.69%),
Severity at Admission	Yes 48 (5.07%)
	No 899 (94.93%),

forest plots (Fig. 2B) to investigate further the direction and extent of the influence of these features. The findings indicated that the risk of whooping cough was significantly and positively correlated with SII (OR: 1.03, 95% CI: 1.01–1.05) and WBC (OR:1.06, 95% CI: 1.01–1.11). This implies that elevated levels of inflammation and white blood cell counts were associated with a higher disease risk. Conversely, the risk of PDW-MPV-RATIO (OR: 0.52, 95% CI: 0.43–0.79) and mean platelet volume (MPV, OR: 0.42, 95% CI: 0.30–0.59) was significantly inversely correlated.

Model performance evaluation and ROC curve analysis

We plotted the ROC curve for the training set and the validation set separately. The random forest performed best in the training set (see S2). Figure 3 depicts the ROC curves of the eight models on the validation set. The ROC curve of the random forest (AUC = 0.98) is close to the upper left corner, indicating high sensitivity and specificity. XGBoost algorithm and GBM have stable classification performance with AUC of 0.97 and 0.95. SVM (AUC = 0.94) performs well but is slightly inferior to the previous models. Logistic regression (AUC = 0.89) has high classification performance and good interpretability. On the other hand, Naive Bayes (AUC = 0.93) and decision trees (AUC = 0.82) performed poorly, with decision trees showing more obvious limitations in challenging tasks. MLP (AUC = 0.50) performed poorly.

Table 2 summarizes the performance of these eight models. In this study, the evaluation results of various models show that random forest and XGBoost algorithm performed best on all key indicators, especially in terms of AUC, accuracy, sensitivity, and F1 score. Specifically, the AUC of random forest was 0.98 and the AUC of XGBoost algorithm was 0.97, both of which showed their extremely high discriminative ability in distinguishing between whooping cough patients and non-whooping cough patients. In addition, the random forest had an accuracy of 0.90 and an F1 score of 0.86, while the XGBoost algorithm had an accuracy of 0.91 and an F1 score of 0.86, indicating that these two models performed well in balancing correct classification and recall.

Among the other models, GBM had an AUC of 0.95, an accuracy of 0.89, and an F1 score of 0.84, which was slightly inferior although close to the performance of random forest and XGBoost algorithm. SVM had AUC of 0.94, an accuracy of 0.89, and an F1 score of 0.85, showing better performance, but overall still lower than random forest and XGBoost algorithm. In contrast, logistic regression and decision trees performed more generally. The AUC of logistic regression was 0.89 and that of decision trees was 0.82, and the F1 scores of the two were 0.78 and 0.54, respectively, indicating that they were inferior to the other more complex models in terms of their ability to distinguish and overall performance.

Overall, random forest and XGBoost algorithm performed better on key indicators such as AUC, accuracy, sensitivity and F1 score, especially when dealing with imbalanced data. These two models excel at balancing classification accuracy and recall rates, and have higher robustness and generalization capabilities. Based on the advantages of random forest in dealing with imbalanced data, low parameter sensitivity and ease of implementation, this study ultimately selected random forest as the main model.

Analyzing calibration curves and decision curves using random forests in the validation set

The calibration curves evaluate the degree to which the model-predicted probability of an event occurring corresponds to the actual likelihood of occurrence. The models performed optimally in the low predictive probability (0–0.2) region, as illustrated in Fig. 4A, with the calibration nearly aligning with the ideal line. The models had adequate calibration at low to moderate risk levels but slightly underestimated the risk of high-risk patients in the region of high predictive probability (> 0.7). In order to ascertain the optimal clinical decision threshold, decision curves evaluated the net benefit at various thresholds. This is particularly relevant to the decision-making process for low and medium-risk patients, as illustrated in Fig. 4B, where the net benefit of the model is considerably more significant than the baseline when the threshold is between 0 and 0.5. The net gain decreases when the threshold surpasses 0.5 but remains above the baseline,

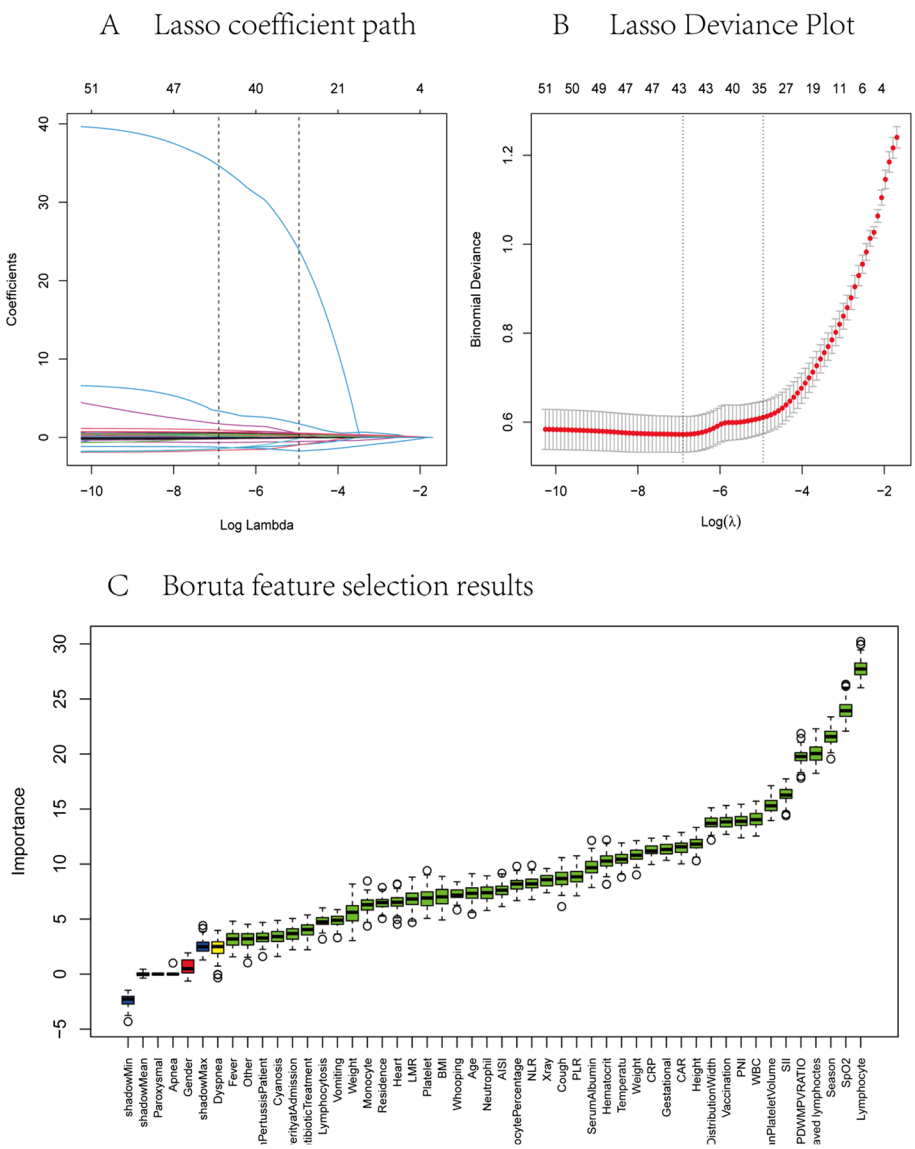

Fig. 1 Lasso regression and Boruta feature selection

suggesting that the model continues to have some clinical value in high-risk patients.

External validation results

ROC curves for the eight models were plotted using external validation data, and the results showed that the AUC of the random forest was 0.97 (S3), which was slightly lower than the AUC of 0.98 on the validation set, but still demonstrated high classification ability. The external validation results further proved that the random forest model also had good generalization performance and was robust on the external validation set. The consistency and robustness of the model on the new dataset was verified, with an accuracy of 89.0%, an F1 score of 0.84, and sensitivities and specificities of 96% and 85%, respectively. Although the performance decreased slightly, the model still showed strong generalization ability (see Table 2).

We performed decision curve and calibration curve analysis on the validation set. The results were similar to those of the validation set. The decision curve showed that the net benefit of the model was better than other strategies at lower thresholds, indicating that it has high clinical value in this range. The calibration curve is close to the ideal line in the medium and low probability ranges, indicating that the model predictions are relatively accurate, but there is some deviation in the high probability region (S4-5).

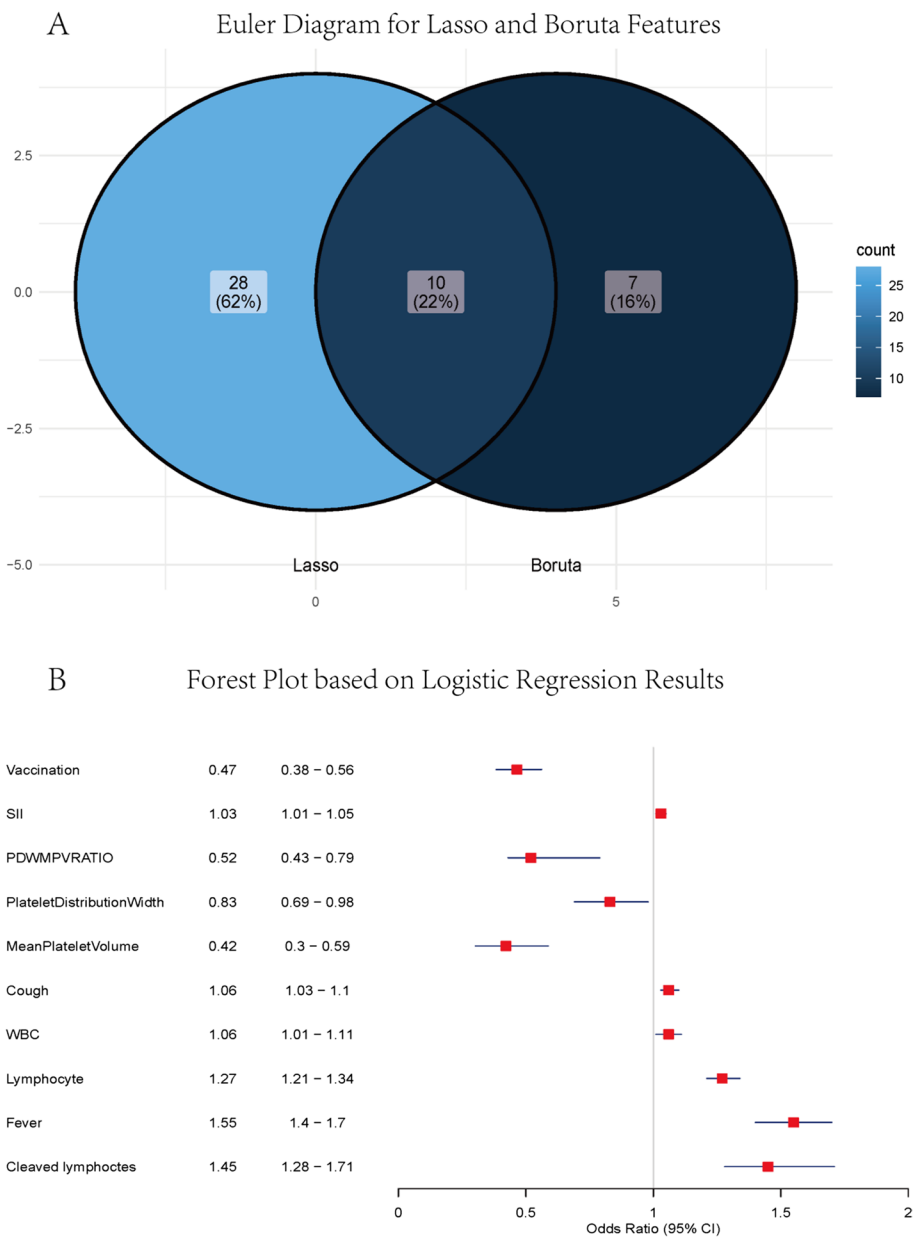

Fig. 2 The Euler diagram and logistic regression

SHAP Scatterplot and online deployment modeling interface using random forests

Figure 5A illustrates the SHAP scatterplot of the model, which illustrates the extent to which each feature contributes to the prediction of pertussis risk. The SHAP value is on the x-axis, while the feature name is on the y-axis. The color of the feature value indicates its magnitude, with blue representing a lesser value and red representing a higher value. SII and WBC SHAP values are widely distributed, suggesting they play a substantial role in predicting a high risk. The SHAP values for Vaccination and PDW-MPV-RATIO are negative, indicating that their high values are associated with a lower risk. SHAP values for Cough and Fever were positive, suggesting that their elevated values were linked to an increased risk. The SHAP plot offers a visual representation of the predictive logic of the model, with ruptured lymphocytes and platelet distribution width having a lesser impact.

We have created an online interface that enables users to input critical clinical characteristics (e.g., SII, PDWMPVRATIO, WBC, number of vaccinations) in order to predict the risk of pertussis in real-time.

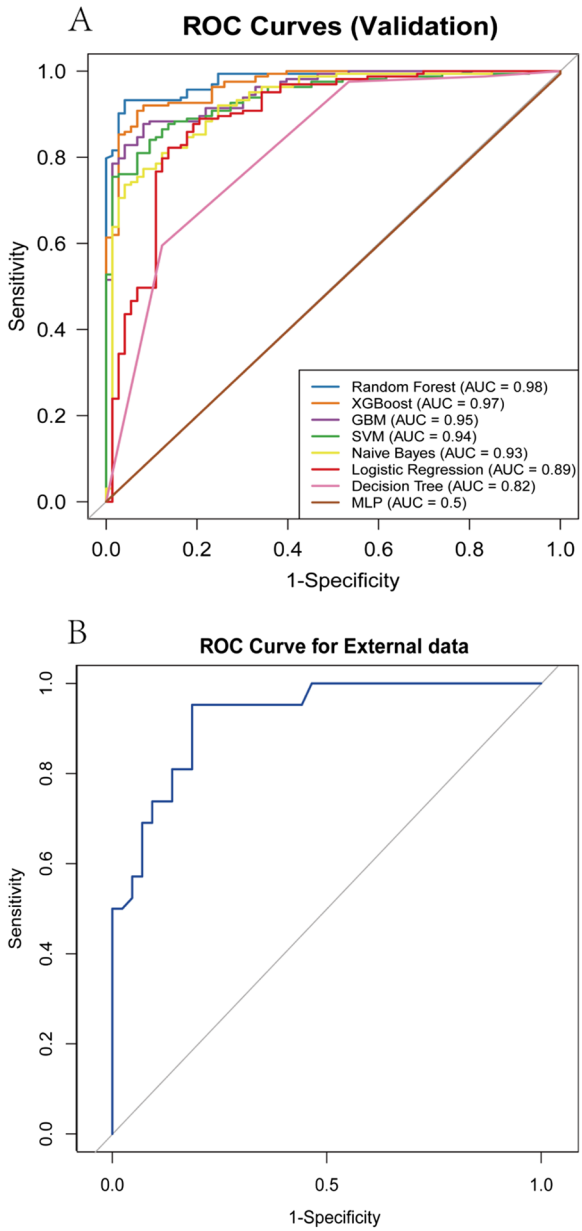

Fig. 3 ROC curve of the eight models and external data

The interface is straightforward and appropriate for researchers and physicians. For instance, the system generates a 94% predicted probability of pertussis and displays the results in a pie chart by inputting four vaccinations, an SII of 816, and a PDW-MPV-RATIO of 1.09. The interface enhances the efficiency of clinical decision-making and offers physicians and researchers convenient instruments for data analysis and diagnosis.

Discussion

This study employed an online platform to clinically implement an efficient pertussis risk prediction model developed using multicenter clinical data [23, 24]. The model exhibited exceptional classification performance in the training set, validation set, and external validation set, thereby illustrating the significant potential of machine learning in predicting respiratory disease risk [25, 26].

Beyond clinical and laboratory indicators, early-life exposures and environmental factors may influence pertussis susceptibility. Prenatal and infant antibiotic exposure can alter immune system development, potentially increasing vulnerability to infections [27]. Likewise, prolonged exposure to air pollution and atmospheric toxins has been linked to respiratory conditions, which may contribute to pertussis risk [28]. While our model currently focuses on clinical and laboratory markers, integrating these broader risk determinants in future iterations may enhance its predictive performance and applicability.

Biological interpretation and feature selection

The Lasso regression and Boruta algorithm were employed to screen critical features, including features like PDW-MPV-RATIO、SII、and WBC, which demonstrated significant roles in predicting pertussis risk [29, 30]. PDW-MPV-RATIO indicates changes in platelet morphology, which implies that platelet activation or vascular pathology changes may be closely associated with the inflammatory response in pertussis [31]. SII, a systemic inflammation indicator, has been extensively employed in the prediction of other infectious diseases, and its validity in pertussis was confirmed in this study [32]. WBC, a traditional marker of inflammation, further supports the significance of inflammation in predicting pertussis risk [33].

The prospective application of cleaved lymphocytes in the early diagnosis of pertussis is attracting attention. Research has demonstrated that children with pertussis have a substantially higher proportion of cleaved lymphocytes and common leukocytes in their peripheral blood. These lymphocytes' lacunar or lobulated nuclear structure is closely linked to the immune responses induced by pertussis toxin [34]. Furthermore, studies have demonstrated that the proportion of slit lymphocytes is higher in neonates under the age of four months, and this cell proportion can serve as a sensitive indicator in the clinical diagnosis of pertussis [35]. In early cases with atypical clinical symptoms, this alteration in cell morphology offers a novel diagnostic approach for pertussis. Additionally, detecting lacunar

Table 2 Performance of the validation set on eight models and external validation

Model	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	F1_score	Kappa
Logistic Regression	0.89	0.85	0.87	0.84	0.71	0.93	0.78	0.67
Random Forest	0.98	0.90	0.97	0.86	0.76	0.98	0.86	0.78
SVM	0.94	0.89	0.93	0.88	0.78	0.96	0.85	0.77
GBM	0.95	0.89	0.89	0.89	0.79	0.95	0.84	0.76
XGBoost algorithm	0.97	0.91	0.93	0.90	0.81	0.96	0.86	0.80
Naive Bayes	0.93	0.85	0.83	0.87	0.74	0.91	0.78	0.68
MLP	0.5	0.86	0.90	0.84	0.72	0.95	0.80	0.7
Decision Tree	0.82	0.80	0.71	0.84	0.67	0.86	0.69	0.54
External data	0.97	0.89	0.96	0.85	0.75	0.97	0.84	0.76

lymphocytes can enhance the accuracy of the diagnosis [36].

These biomarkers improve the predictive potential of the model; however, additional validation is necessary to guarantee their performance in a broader clinical context and to ensure their applicability in various age groups or immune statuses [37, 38].

Influence of vaccine efficacy and immune evasion

Pertussis recurrence has become a concern in certain regions in recent years due to vaccine efficacy's potential immune escape and attenuation [39, 40]. The results of this study indicate that vaccination remains an effective method of prevention, as there is a substantial negative correlation between vaccination and the risk of pertussis [41, 42]. Nevertheless, the protective effect of vaccines may diminish over time, and immune escape may impact risk prediction [43]. Consequently, in the future, a time-after-vaccination variable could be implemented to evaluate fluctuations in vaccine efficacy dynamically, and the predicted outcomes could be further examined about the effects of various vaccine types [44].

Clinical and explanatory acceptance of machine learning models

This study improved the interpretability of the model through SHAP analysis, which allowed clinicians to gain a more comprehensive understanding of the prediction results [45, 46]. By incorporating the SHAP results into an online tool and utilizing a graphical representation of each feature's contribution to the prediction results, physicians' confidence in the model is increased [29]. This explanatory enhancement facilitates physicians' comprehension of intricate machine-learning models and encourages their pervasive implementation in clinical practice [47]. Simplified user interfaces and intuitive explanations are critical in the community of clinicians from non-technical backgrounds [48].

The clinical importance of calibration and decision curves

The calibration curve demonstrated that the model was more accurate in predicting low-risk patients, which could assist physicians in avoiding superfluous interventions and reducing the waste of clinical resources [49]. The decision curve demonstrated that the model's net benefit was enhanced under various thresholds for high-risk patients despite a modest calibration bias. This was particularly advantageous for screening low and medium-risk patients [50]. This implies that the model can be employed to enhance the accuracy of clinical decision-making, optimize the management of low- and medium-risk patients, and identify high-risk patients [51].

Public health consequences and application perspectives

The online model studied in this study provides a pertussis surveillance and control instrument that is effective for public health departments [37]. The model can identify high-risk populations in real time and provide data support for public health interventions, particularly in under-vaccinated or outbreak areas, by integrating into regional or national infectious disease prevention and control platforms [52]. The study could expand the model's implementation by combining the risk assessment of other contagious diseases to create a future comprehensive public health defense and control platform [32].

Limitations

Although the results of this investigation are encouraging, there are still some constraints:

1. Data source constraints: the model's global generalization is restricted because the training and validation data are primarily sourced from specific regions, which may have distinct demographic and geographic characteristics [53]. Further validation with datasets from the additional areas is required to guar-

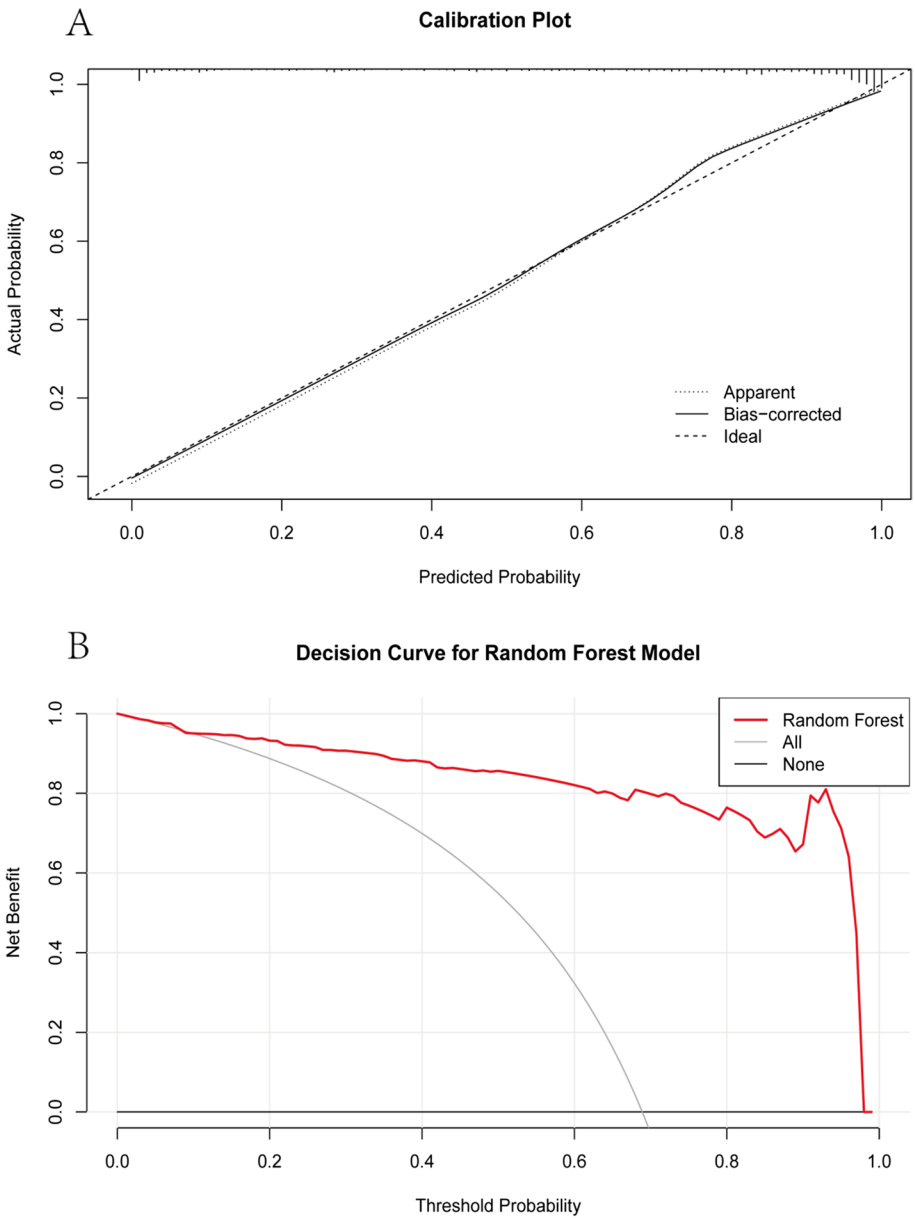

Fig. 4 Calibration curve and decision curve

antee the model's general applicability, as epidemiological traits may differ from region to region [54].

2. Model interpretability and clinical acceptance: Despite SHAP analysis enhancing model interpretability, clinicians may continue to harbor reservations regarding the "black box" nature of machine learning models [55]. The issue of improving models' transparency without compromising efficacy still needs to be solved [56].
3. The model's performance should be further validated in the future by a more diverse dataset, as the scale and diversity of external validation were relatively limited, even though the model was validated for stability by external data [57].
4. The study could not capture patient changes during the disease due to the absence of dynamic data support based on a static dataset. In the future, the model's predictive capabilities can be enhanced by incorporating dynamic data or time series analysis to more effectively adapt to real-time data in the clinic [58].

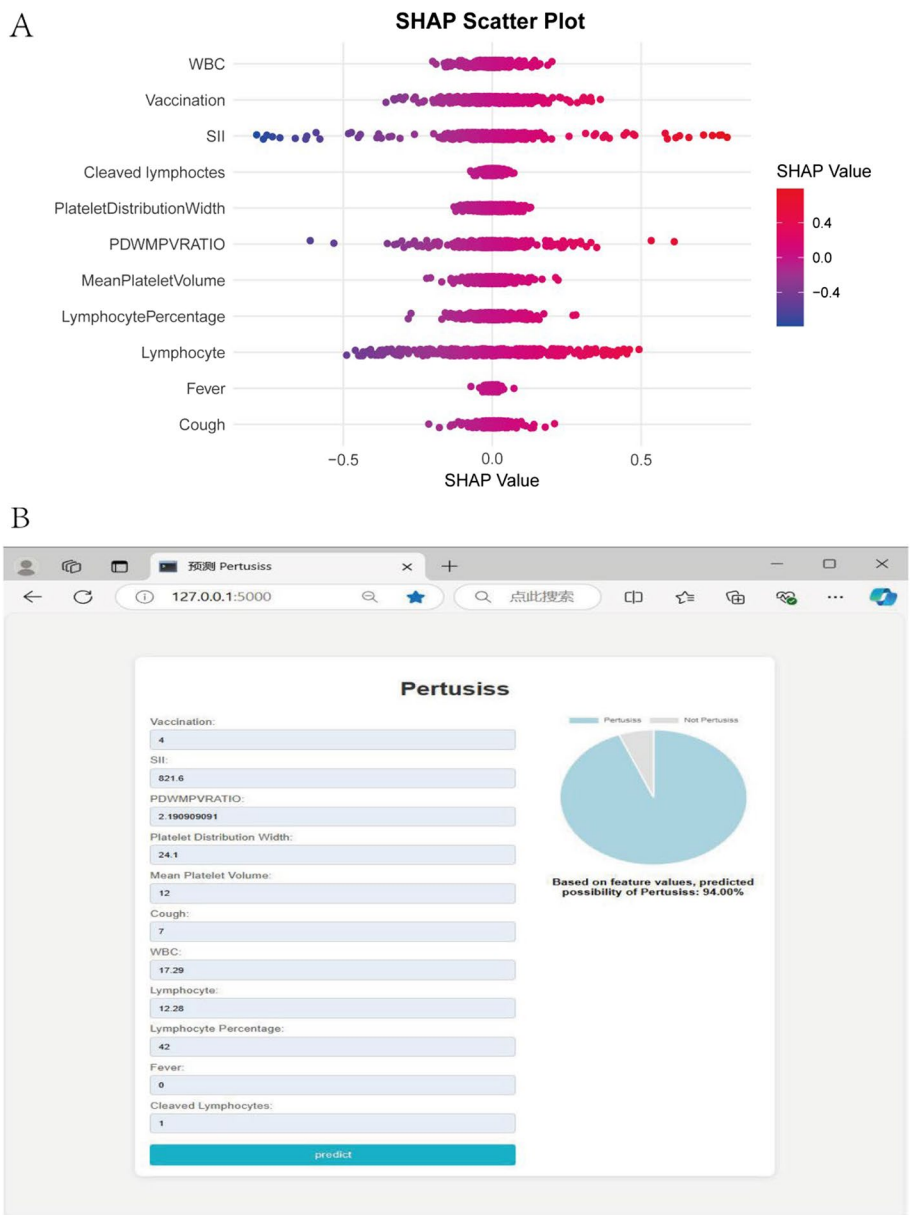

Fig. 5 SHAP scatterplot and online deployment modeling

Future studies

To verify the model's global applicability, future research should broaden the data sources to encompass patient data from various regions, ethnicities, and age groups [33]. Furthermore, the model's real-time prediction capability will be enhanced by incorporating dynamic features or time-series data to better its ability to capture disease progression [59]. The model's efficacy in real-world clinical settings will be further validated through prospective clinical studies, ensuring seamless integration into clinical workflows and additional support to physicians and public health policymakers [60, 61].

Conclusion

This study employed multicenter data to develop a machine-learning model to facilitate early pertussis detection. The model demonstrated strong generalizability and predictive performance, with successful online deployment for clinical use. The model offers robust support for the early prediction and public health prevention and control of pertussis despite the scope of validation

and the limitations of data sources. Integrating dynamic data and broadening the scope of data sources is advisable to enhance the model's accuracy and applicability.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-025-10797-7.

Supplementary Material 1

Acknowledgements
We would like to express our sincere gratitude to Juan Xie for her valuable contributions to revising this manuscript. We also thank all the hospital staff involved in data collection and the study participants for their cooperation. Additionally, we are deeply grateful to all the patients who participated in this study for their invaluable contribution. We also extend our heartfelt thanks to the Kunming Children's Hospital for their support and collaboration throughout the study.

Clinical trial number
Not applicable.

Authors' contributions
J.X. and R.M. wrote the main manuscript text. Y.F. and Y.Q. prepared Figs. 1- 3. H.Z. and X.T. conducted the data analysis. W.C. and C.L. reviewed the manuscript. L.T. and L.C. provided critical revisions and feedback. All authors reviewed and approved the final manuscript.

Funding
2023 Kunming Health Research Project (Project No. 2023–06-01–038).

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Our study was according to the ethical guidelines of the Helsinki Declaration and was approved by the Human Ethics Committee of the Kunming Children Hospital. Parents or legal guardians of all participants under the age of 16 have provided informed consent. All participants aged 16 years or above have also provided informed consent.

Consent for publication
Written informed consent was obtained from the patients for publication.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Anesthesiology, Kunming Children'S Hospital, Kunming City, Yunnan Province, China. [2]Department of Cardiac Surgery, Fuwai Yunnan Hospital, Chinese Academy of Medical Sciences/Affiliated Cardiovascular Hospital of Kunming Medical University, Kunming City, Yunnan Province, China. [3]Comprehensive Pediatrics, Wenshan Maternal and Child Health Care Hospital, Wenshan City, Yunnan Province, China. [4]Comprehensive Pediatrics and Neonatology, Chuxiong Yi Autonomous Prefecture People's Hospital, Chuxiong City, Yunnan Province, China. [5]Pediatric Respiratory Department, Qujing Maternal and Child Health Hospital, Qujing City, Yunnan Province, China. [6]Department of Pediatrics, Kaiyuan People's Hospital, Kaiyuan, China. [7]Department of Pediatrics and Emergency, Yuxi Children'S Hospital, Yuxi City, Yunnan Province, China. [8]Comprehensive Pediatrics & Pulmonary and Critical Care Medicine, Baoshan People's Hospital, Baoshan City, Yunnan Province, China. [9]Department of Respiratory Medicine Kunming Children'S Hospital, Kunming City, Yunnan Province, China. [10]Comprehensive Pediatrics & Pulmonary and Critical Care Medicine, Kunming Children'S Hospital, Yunnan Province, Shulin Street 28, Kunming City, Yunnan Province 650000, China.

Received: 4 December 2024 Accepted: 13 March 2025
Published online: 27 March 2025

References

1. Kohberger R, Jemiolo D, Noriega F. Prediction of pertussis vaccine efficacy using a correlates of protection model. Vaccine. 2008;26(27–28):3516–21.
2. Queenan A, Dowling DJ, Cheng W, Faé K, Fernandez J, et al. Increasing FIM2/3 antigen-content improves efficacy of Bordetella pertussis vaccines in mice in vivo. Vaccine. 2019;37(1):80–9.
3. Safan M, Barley K, Elhaddad MM, Darwish MA, Saker S. Mathematical analysis of an SIVRWS model for pertussis with waning and naturally boosted immunity. Symmetry. 2022;14:2288.
4. Guinto-Ocampo H, Bennett J, Attia M. Predicting pertussis in infants. Pediatr Emerg Care. 2007;24:16–20.
5. Daluwatte C, Dvaretskaya M, Ekhtiari S, Hayat P, Montmerle M, Mathur S, Macina D. Development of an algorithm for finding pertussis episodes in a population-based electronic health record database. Hum Vaccin Immunother. 2023;19:2209455.
6. Witt M, Arias L, Katz PH, Truong E, Witt D. Reduced risk of pertussis among persons ever vaccinated with whole cell pertussis vaccine. Clin Infect Dis. 2013;56(9):1248–54.
7. Le T, Cherry J, Chang S, Knoll M, Lee M, Barenkamp S, et al. Immune responses and antibody decay after immunization with an acellular pertussis vaccine. J Infect Dis. 2004;190(3):535–44.
8. Dabrera G, Amirthalingam G, Andrews N, Campbell H, Ribeiro S, Kara E, et al. A case-control study to estimate the effectiveness of maternal pertussis vaccination in protecting newborn infants. Clin Infect Dis. 2015;60(3):333–7.
9. Fulton TR, Phadke VK, Orenstein W, Hinman A, Johnson W, Omer S. Protective effect of contemporary pertussis vaccines: a systematic review and meta-analysis. Clin Infect Dis. 2016;62(9):1100–10.
10. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak. 2019;19(1):281.
11. Pudjihartono N, Fadason T, Kempa-Liehr A, O'Sullivan J. A review of feature selection methods for machine learning-based disease risk prediction. Front Bioinforma. 2022;2:927312.
12. Bertini A, Salas R, Chabert S, Sobrevia L, Pardo F. Using machine learning to predict complications in pregnancy: a systematic review. Front Bioeng Biotechnol. 2022;9:780389.
13. Visumathi A, Velagapudi A, Reddy R, Anil Kumar P. Multi-Disease Prediction Using Machine Learning Algorithm. In: International Journal for Research in Applied Science and Engineering Technology Conference. Chennai: 2023.
14. Banerjee A, Chen S, Fatemifar G, Zeina M, Lumbers RT, Mielke J, Gill S, Kotecha D, Freitag D, Denaxas S, Hemingway H. Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility. BMC Med. 2021;19:1-4.
15. Grampurohit S, Sagarnal C. Disease prediction using machine learning algorithms. International Conference for Emerging Technology (INCET). 2020;2020:1–7.
16. Kohli P, Arora S. Application of Machine Learning in Disease Prediction. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA). 2018.
17. Daluwatte C, Dvaretskaya M, Ekhtiari S, Hayat P, Montmerle M, Mathur S, Macina D. Development of an algorithm for finding pertussis episodes in a population-based electronic health record database. Hum Vaccin Immunother. 2023;19:19.
18. Tozzi A, Gesualdo F, Rizzo C, Carloni E, Russo L, Campagna I, Villani A, Reale A, Concato C, Linardos G, Pandolfi E. A data driven clinical algorithm for differential diagnosis of pertussis and other respiratory infections in infants. PLoS One. 2020;15(7):e0236041.
19. Karunya K, M Nivetha, Selvan HT, Janani GS. Anatomization of Respiratory Diseases Using Machine Learning. Recent Trends in Artificial Intelligence & its Applications. 2023;2(3):17-27.

20. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access. 2019;7:81542–54.
21. Thiry N, Boutriau D, Bogaerts H, Beutels P, Reinert P. Impact of childhood vaccination strategies on pertussis epidemiology: a model comparison study. Vaccine. 2012;30(28):4251–60.
22. Castagnini LA, Heininger U. Pertussis resurgence: what are the solutions? Expert Rev Respir Med. 2015;9(2):175–7.
23. McCord-De Iaco K, Gesualdo F, Pandolfi E, Croci I, Tozzi A. Machine learning clinical decision support systems for surveillance: a case study on pertussis and RSV in children. Front Pediatr. 2023;11:1112074.
24. Jürgen D, Sabrina K. Assessing the strengths and limitations of LIME for model interpretability. Int J Data Sci Anal. 2020;7:479. https://doi.org/10.7717/peerj-cs.479.
25. Tideman L, Migas L, Djambazova KV, et al. Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations. Anal Chim Acta. 2021;1177:338522.
26. Schilling V, Beyerlein P, Chien J. A bioinformatics analysis of ovarian cancer data using machine learning. Algorithms. 2023;16(7):330.
27. Lee SH, Kim BN, Kim JW, et al. Prenatal and infant exposure to antibiotics and subsequent risk of neuropsychiatric disorders in children: a nationwide birth cohort study in South Korea. JAMA Pediatr. 2023;177(5):485–94.
28. Wang X, Li Y, Zhao H, et al. Atmospheric environment and persistence of pediatric asthma: a population-based cohort study. Environ Health Perspect. 2023;131(3): 037005.
29. Frénard C, Blanchet K, Lecerf P, et al. Machine learning algorithm to predict response to immunotherapy in real-life settings for patients with advanced melanoma. Eur J Dermatol. 2023;33(2):75–80.
30. Parisi L, RaviChandran N, Manaog ML. A novel hybrid algorithm for aiding prediction of prognosis in patients with hepatitis. Neural Comput Appl. 2019;32:3839–52. https://doi.org/10.1007/s00521-019-04050-x.
31. Chen X, Li Y, Li X, et al. An interpretable machine learning prognostic system for locoregionally advanced nasopharyngeal carcinoma based on tumor burden features. Oral Oncol. 2021;118: 105335.
32. Yagin B, Yagin F, Colak C, et al. Cancer metastasis prediction and genomic biomarker identification through machine learning and eXplainable artificial intelligence in breast cancer research. Diagnostics. 2023;13(21):3314.
33. Korfiatis P, Kline T, Coufalová L, et al. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas. Med Phys. 2016;43(6):2835–44.
34. Xu X, Zhang J. Clinical value of cleaved lymphocytes in assisting the diagnosis of pertussis in children. Zhongguo Dang Dai Er Ke Za Zhi. 2020;22(9):996–1000.
35. Zhang R, Wang H, Li C, et al. Utility of cleaved lymphocytes from peripheral blood smear in the diagnosis of pertussis. Int J Lab Hematol. 2020;43:1–10.
36. Kubic V, Kubic P, Brunning R. The morphologic and immunophenotypic assessment of the lymphocytosis accompanying Bordetella pertussis infection. Am J Clin Pathol. 1991;95(6):809–15.
37. Ikemura K, Bellin E, Yagi Y, et al. Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study. J Med Internet Res. 2021;23(2):e23458. https://doi.org/10.2196/23458.
38. Wang J, Gao W, Lu M, et al. Development of an interpretable machine learning model for Ki-67 prediction in breast cancer using intratumoral and peritumoral ultrasound radiomics features. Front Oncol. 2023;13:1290313. https://doi.org/10.3389/fonc.2023.1290313.
39. Hathaway Q, Roth SM, Pinti M, et al. Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. Cardiovasc Diabetol. 2019;18:87.
40. Burnham-Marusich AR, Olsen RK, Scarbrough J, et al. Tracheal colonization factor A (TcfA) is a biomarker for rapid and specific detection of Bordetella pertussis. Sci Rep. 2020;10:72092–6.
41. Markey K, Douglas-Bardsley A, Hockley J, Le Tallec D, Costanzo A. Calibration of pertussis toxin BRP batch 1 in a standardised CHO cell-based clustering assay. Pharmeur Bio Sci Notes. 2018;2018:112-23.
42. Pramono RX, Imtiaz SA, Rodriguez-Villegas E. A Cough-Based Algorithm for Automatic Diagnosis of Pertussis. PLoS One. 2016;11(9):e0162128.
43. Taneja I, Reddy B, Damhorst G, et al. Combining biomarkers with EMR data to identify patients in different phases of sepsis. Sci Rep. 2017;7:09766–71.
44. Dou M, Macias N, Shen F, et al. Rapid and accurate diagnosis of the respiratory disease pertussis on a point-of-care biochip. EClinicalMedicine. 2019;8:72–7.
45. Kamal SA, Yin C, Qian B, Zhang P. An interpretable risk prediction model for healthcare with pattern attention. BMC Med Inform Decis Mak. 2020;20(Suppl 11):307.
46. Tallarida RJ. Quantitative methods for assessing drug synergism. Genes Cancer. 2011;2(11):1003–8. https://doi.org/10.1016/j.drudis.2011.04.011.
47. Sato M, Morimoto K, Kajihara S, et al. Machine-learning approach for the development of a novel predictive model for the diagnosis of hepatocellular carcinoma. Sci Rep. 2019;9(9):1–9.
48. Enhancing trust and interpretability of complex machine learning models using SHAP explanations. J Comput Aided Mol Des. 2021;35:712–728.
49. Saeidpour A, Bansal S, Rohani P. Dissecting recurrent waves of pertussis across the boroughs of London. PLoS Comput Biol. 2022;18(4):e1009898.
50. Siah KW, Khozin S, Wong CH, Lo A. Machine-learning and stochastic tumor growth models for predicting outcomes in patients with advanced non-small-cell lung cancer. JCO Clin Cancer Inform. 2019;3:1–11.
51. Weaver KL, Blackwood CB, Horspool AM, Pyles GM, Sen-Kilic E, Grayson EM, Huckaby AB, Witt WT, DeJong MA, Wolf MA, Damron FH, Barbier M. Long-Term Analysis of Pertussis Vaccine Immunity to Identify Potential Markers of Vaccine-Induced Memory Associated With Whole Cell But Not Acellular Pertussis Immunization in Mice. Front Immunol. 2022;13:838504.
52. Fan R, Qin W, Zhang H, et al. Machine learning in the prediction of cardiac surgery associated acute kidney injury with early postoperative biomarkers. Front Surg. 2023;10: 1048431. https://doi.org/10.3389/fsurg.2023.1048431.
53. Wang F, Xu C, Chen W, Duan S. A Glycolysis Gene Methylation Prediction Model Based on Explainable Machine Learning for Alzheimer's Disease. SSRN Electronic Journal. 2021.
54. Gao M, Huang S, Pan X, et al. Machine-learning based radiomics predicting tumor grades and expression of multiple pathologic biomarkers in gliomas. Front Oncol. 2020;10:1676.
55. Li W, Zhu L, Li K, et al. Machine learning-assisted dual-marker detection in serum small extracellular vesicles for the diagnosis and prognosis prediction of non-small cell lung cancer. Nanomaterials. 2022;12(5): 809.
56. Parisi L, RaviChandran N, Manaog ML. A novel hybrid algorithm for aiding prediction of prognosis in patients with hepatitis. Neural Comput Appl. 2019;32:3839–52.
57. Ma T, Zhang Y, Zhao M, et al. A machine learning-based radiomics model for prediction of tumor mutation burden in gastric cancer. Front Genet. 2023;14: 1283090.
58. Jia M, Wu Y, Xiang C, Fang Y. Predicting Alzheimer's disease with interpretable machine learning. Dement Geriatr Cogn Disord. 2023;52(4):249–57.
59. Konerman MA, Beste LA, Van T, Liu B, Zhang X, Zhu J, Saini SD, Su GL, Nallamothu BK, Ioannou GN, Waljee AK. Machine learning models to predict disease progression among veterans with hepatitis C virus. PLoS One. 2019;14(1):e0208141.
60. Coudeville L, Van Rie A, Getsios D, Caro JJ, Crépey P, Nguyen VH. Adult vaccination strategies for the control of pertussis in the United States: an economic evaluation including the dynamic population effects. PLoS One. 2009;4(7):e6284.
61. Esposito S, Stefanelli P, Fry NK, Fedele G, He Q, Paterson P, Tan T, Knuf M, Rodrigo C, Olivier CW, Flanagan KL, Hung I, Lutsar I, Edwards K, O'Ryan M, Principi N. Pertussis Prevention: Reasons for Resurgence, and Differences in the Current Acellular Pertussis Vaccines. Front Immunol. 2019;10:1344.

Publisher's Note