Diagnostic Accuracy Study

Medicine

OPEN

Machine-learning-algorithms-based diagnostic model for influenza A in children

Qian Zeng, MD[a,b], Chun Yang, BS[a,b], Yurong Li, BS[a,b], Xinran Geng, MD[c], Xin Lv, PhD[a,b,*]

Abstract

Background: At present, nucleic acid testing is the gold standard for diagnosing influenza A, however, this method is expensive, time-consuming, and unsuitable for promotion and use in grassroots hospitals. This study aimed to establish a diagnostic model that could accurately, quickly, and simply distinguish between influenza A and influenza like diseases.

Methods: Patients with influenza-like symptoms were recruited between December 2019 and August 2023 at the Children's Hospital Affiliated to Shandong University and basic information, nasopharyngeal swab and blood routine test data were included. Computer algorithms including random forest, GBDT, XGBoost and logistic regression (LR) were used to create the diagnostic model, and their performance was evaluated using the validation data sets.

Results: A total of 4188 children with influenza-like symptoms were enrolled, of which 1992 were nucleic acid test positive and 2196 were matched negative. The diagnostic models based on the random forest, GBDT, XGBoost and logistic regression algorithms had AUC values of 0.835,0.872,0.867 and 0.784, respectively. The top 5 important features were lymphocyte (LYM) count, age, serum amyloid A (SAA), white blood cells (WBC) count and platelet-to-lymphocyte ratio (PLR). GBDT model had the best performance, the sensitivity and specificity were 77.23% and 80.29%, respectively.

Conclusions: A computer algorithm diagnosis model of influenza A in children based on blood routine test data was established, which could identify children with influenza A more accurately in the early stage, and was easy to popularize.

Abbreviations: AUC = area under the curve, CRP = C-reactive protein, LC = product of lymphocytes and C-reactive protein, LCR = lymphocyte-to-C-reactive protein ratio, LMR = lymphocyte-to-monocyte ratio, LR = logistic regression, LYM = lymphocyte, MON = Monocyte count, MPV = mean platelet volume, NEU = neutrophil count, NLR = neutrophil-to-lymphocyte ratio, PLR = platelet-to-lymphocyte ratio, PLT = platelet, ROC = receiver operator characteristic, RT-PCR = reverse transcriptase-polymerase chain reaction, SAA = serum amyloid A, WBC = white blood cells.

Keywords: children, diagnostic model, GBDT, influenza A, machine-learning-algorithms

1. Introduction

Influenza is an acute respiratory infectious disease caused by influenza viruses belonging to the *Orthomyxoviridae* family of single-stranded RNA viruses. It is the most common cause of respiratory infections in children annually with significant disease burden, high proportion of hospitalization and substantial morbidity and mortality worldwide.[1–4] The clinical presentation of influenza in children shares many nonspecific upper respiratory features including cough, fever, headache, general malaise, and myalgia, making it difficult to distinguish influenza from other respiratory viral pathogens.[5–7] Therefore, early diagnosis is essential to reduce children suffering, treatment costs and the incidence of complications.

Nowadays the most ideal, sensitive and specific diagnostic detection method of influenza is reverse transcriptase-polymerase chain reaction (RT-PCR) of nasopharyngeal swab specimens in tertiary hospital.[8] Although it is fast and sensitive, it is relatively expensive and not suitable for large-scale promotion in community hospitals. In summary, we urgently need a common, cheap, and versatile test indicator to screen for influenza viruses.

In the era of big data, medical database and computer algorithm can be applied to influenza epidemic trend prediction, disease outcome, prognosis judgment and other aspects. In the field of assisted diagnosis of diseases, the algorithm can complete complex analysis tasks in a very short time after learning electronic medical records and other relevant information, and feedback the best classification pattern results according to the input information, helping physicians to improve the accuracy and efficiency of patient diagnosis. This method has been widely used in medical research of various diseases.[9,10]

The authors have no funding and conflicts of interest to disclose.

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

[a] Clinical Laboratory, Children's Hospital Affiliated to Shandong University, Jinan, China, [b] Clinical Laboratory, Jinan Children's Hospital, Jinan, China, [c] Maternity & Child Care Center of Dezhou, China.

**Correspondence: Xin Lv, Clinical Laboratory, Children's Hospital Affiliated to Shandong University, 23976 Jing-Shi Road, Jinan 250022, Shandong Province, PR China (e-mail: etyyjyklvxin@163.com).*

How to cite this article: Zeng Q, Yang C, Li Y, Geng X, Lv X. Machine-learning-algorithms-based diagnostic model for influenza A in children. Medicine 2023;102:48(e36406).

Received: 3 September 2023 / Received in final form: 8 November 2023 / Accepted: 10 November 2023

http://dx.doi.org/10.1097/MD.0000000000036406

However, computer algorithm technology is rarely used in the research of auxiliary diagnosis of influenza in children at present. In this paper, appropriate clinical information and laboratory indicators will be selected in combination with computer algorithm to establish a diagnostic model for influenza A in children.

Complete blood count is the first choice in children with influenza symptoms, which can be carried out in hospitals of different grade and easy to perform. Moreover, recent studies have also shown that hematological indexes based on the analysis of blood routine and classification are of great clinical significance as screening markers for the diagnosis of influenza, such as neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), lymphocyte-to-monocyte ratio (LMR) and mean platelet volume-to-platelet ratio (MPV/PLT).[11–14] In addition, many studies have confirmed that C-reactive protein (CRP) and Serum amyloid A (SAA) changes when influenza invades the human body, especially in severely ill patients.[15–17] They are acute-phase proteins downstream of the pro-inflammatory cytokines released during infection progress, which play an intermediate role in the relationship between serum cytokine oversynthesis and influenza.[18,19]

Therefore, this study introduced machine algorithm for the first time, applied common clinical influenza screening indicators to establish a comprehensive, accurate and rapid childhood influenza A diagnosis model, and verified its diagnostic efficiency.

2. Materials and Methods

2.1. Study population

This retrospective study was conducted in clinical laboratory department of Children Hospital Affiliated to Shandong University from December 2019 to August 2023.The study was approved by the Ethical Board of Children's Hospital Affiliated to Shandong University. Children with influenza-like symptoms who underwent routine blood tests and the RT-PCR for influenza virus at the first visit were enrolled in this study. Influenza-like symptoms were defined according to the National Health Commission "Influenza Diagnosis and Treatment Plan (2020 Edition)" guidelines as follows: fever, headache, myalgia, chills, cough, respiratory tractinfection, sore throat and other systemic symptoms. Exclusion criteria were: 1, history of systemic chronic diseases, autoimmune disease, malignancy; 2, concurrent bacterial illness (bacteremia, bacterial pneumonia) or other viral infections; 3, long-term use of immunosuppressive medications. Finally, a total of 4188 eligible children were enrolled, of which 1992 were positive for influenza A. To match the positive group, 2196 negative children were randomly enrolled, and their basic information was shown in Table 1.

2.2. Research feature selection

According to clinical experience and hospital database data records, this study recruited a total of 16 features in 3 categories: General data indicators: gender and age. Blood cell test results: white blood cell count (WBC), platelet count (PLT), neutrophil count (NEU), lymphocyte count (LYM), Monocyte count (MON), Mean Platelet Volume (MPV), CRP, SAA. Calculation results: NEU/LYM, PLT/LYM, LYM/MON, LYM/CRP, LYM*CRP, CRP/SAA.

2.3. Data preprocessing

The collection and processing of the results was carried out by different people, and the grouped information was blind to them. Data processing involved 3 steps: Fill in missing value, process data beyond the linear range and numeralize categorical variables. Missing values. The missing values were confirmed by medical experts and supplemented by the average value. Data outside the linear range. All data outside the linear range was represented as the lowest or highest measured value, which was only seen in CRP and SAA. Categorical variables. Categorical variables such as gender, negative or positive are converted to numbers 1 or 0; The age of the children was converted into numbers in months.

Table 1

Statistical results of measurement data of 2 groups. (Median (interquartile range)).

	FluA positive (1992)	FluA negative (2196)	*P* value
Gender	1152/840 (Boys/Girls)	1270/926 (Boys/Girls)	.999
Age	60 (42–79)	38 (16–61)	<.000
WBC	6.15 (4.67–8.08)	7.75 (5.51–10.72)	<.000
LYM	1.42 (0.92–2.23)	2.51 (1.69–3.65)	<.000
NEU	3.86 (2.56–5.48)	3.92 (2.32–6.51)	.082
MON	0.54 (0.39–0.72)	0.61 (0.42–0.88)	<.000
MPV	8.9 (8.4–9.5)	8.9 (8.3–9.5)	<.000
PLT	229 (192–270)	263 (211–325)	<.000
SAA	26.98 (10.61–56.89)	21.93 (6.11–65.78)	.001#
CRP	3.56 (0.87–8.36)	3.86 (0.50–11.07)	.055
NLR	2.73 (1.42–4.93)	1.53 (0.83–2.96)	<.000
PLR	158.69 (103.90–250.00)	101.42 (73.61–148.68)	<.000
LMR	2.59 (1.74–4.17)	4.06 (2.57–6.46)	<.000
LCR	0.41 (0.16–1.48)	0.64 (0.20–3.60)	<.000
LC	4.37 (1.39–12.51)	8.1 (2.01–27.23)	<.000
CRP/SAA	0.13 (0.07–0.23)	0.12 (0.07–0.28)	.067#

LC = LYM*CRP, LCR = LYM/CRP, LMR = LYM/MON, NLR = NEU/LYM, PLR = PLT/LYM.
# Comparison results before missing values are filled.

2.4. Model construction

The data set was randomly divided into training data set and validation data set according to the ratio of 6:4, and 10-fold cross-validation was used to test the accuracy of the algorithm. In the training data set, random forest, GBDT, XGBoost and logistic regression algorithms were used to construct children influenza A diagnosis models respectively using SPSSPRO software. The selection of model parameters and hyperparameters (n_estimators, max_depth, max_features, min_samples_leaf, min_samples_split, learning_rate, loss and so on) was set using SPSSPRO software according to the number and type of data.

2.5. Model algorithm

2.5.1. Random Forest. Random forest is a classifier containing many decision trees, which can be used for both classification and regression problems and dimension reduction problems. It also has good tolerance for outliers and noise, and has better prediction and classification performance than decision trees.

2.5.2. GBDT model. GBDT is one of the best algorithms to fit the real distribution in machine learning algorithms. By setting a threshold, greater than the threshold is a positive example, and vice versa is a negative example. It is suitable for classification and regression problems, and can screen features, and is one of the most commonly used models in medical diagnosis research.

2.5.3. XGBoost model. XGBoost is a scalable tree boosting algorithm that is widely used in data science. XGBoost can be said to be an improved version of GBDT, which can train models faster and more efficiently.

2.5.4. LR model. LR is a generalized linear regression classification model commonly used in disease prediction research. By inputting the characteristic properties of unknown samples, the probability that the samples belong to a certain class can be calculated.

2.6. Model evaluation

The validation data set was used for the validation and evaluation of the model. Receiver operator characteristic curve (ROC curves), area under the curve (AUC), sensitivity, specificity, negative predictive value and positive predictive value were used to evaluate the diagnostic efficacy of each model. The specific research flow diagram was shown in Figure 1.

2.7. Statistical analysis

All statistical analyses were evaluated by using SPSS 17.0 software. Quantitative data consistent with normal distribution were expressed as mean ± standard, and inter-group comparison was performed by Student *t* test. Quantitative data that did not conform to the normal distribution were expressed as the median (interquartile range) and compared between groups using the Wilcoxon rank sum test. Qualitative data were presented in example (n/n), and Chi-square test was used for inter-group comparison. Significant differences were accepted at $P < .05$.

3. Results

3.1. Patient characteristics and hematological parameters in 2 groups

This study recruited 4188 influenza-like children eventually and divided them into Flu A + group and Flu- group based on the results of RT-PCR. There was no significant difference in gender composition between the 2 groups. However, there were significant differences in age distribution, as shown in Table 1. The results of complete blood count and related hematological parameters in Flu A + and Flu- groups were presented in Table 1. There were significantly different in all other indicators between the 2 groups, except for NEU count, CRP and CRP/SAA. Compared with the Flu- group, the A + group had lower WBC count, LYM count, MON count, PLT count, LMR, LCR and LC values, and higher SAA, NLR, and PLR values.

3.2. Diagnostic performance of the model

In this study, evaluation indexes such as AUC, sensitivity, specificity, negative predictive value and positive predictive value were selected to evaluate the results of the validation data set of 4 machine learning modeling algorithms. Specific experimental results were shown in Table 2, heatmap and ROC curves of each algorithm were shown in Figures 2 and 3. As can be seen from the results in Table 2, for the experimental data set of this study, the GBDT algorithm had the best prediction results, with an AUC value of 0.872, sensitivity and specificity of 77.23%, 80.29%, respectively. The AUC values of random forest, XGBoost and logistic regression algorithms were 0.835, 0.867 and 0.784, respectively. The specificity was 75.13%, 79.66% and 74.53%, respectively.

3.3. Model feature importance

Random forest, GBDT, XGBoost and logistic regression machine learning algorithms were used to model. Figure 4 showed the feature variable importance results of various algorithms. Combined with the feature importance of the 4 modeling algorithms, it could be concluded that variables such as LYM count, age, SAA, WBC and PLR might play an important role in predicting influenza A.

4. Discussions

The influenza viruses cause epidemics in humans. It is estimated that 5% to 10% of adults and 20% to 30% of children are infected and causing 3 to 5 million severe cases besides 250,000 to 500,000 deaths annually.[20,21] In the seasonal influenza epidemics children have the highest attack rates and be more vulnerable with severe complications.[22,23] In summary, early diagnosis and treatment of influenza is the main point to reducing severe illness and complications.[18]

In the era of big data, machine algorithm model as an analytical tool, cannot be ignored in the analysis of precise diagnostic models for influenza A. This article uses 4 machine algorithms models for auxiliary diagnosis of influenza and compares the results of them. Among them, the GBDT model has the best performance in distinguishing between influenza A and influenza like diseases, with an AUC value of up to 0.872, followed by XGBoost and random forest model, and the model has good sensitivity and specificity (sensitivity: 77.23%; specificity: 80.29%). In terms of the diagnostic performance of influenza A and influenza like illness, our model is significantly outperformed than

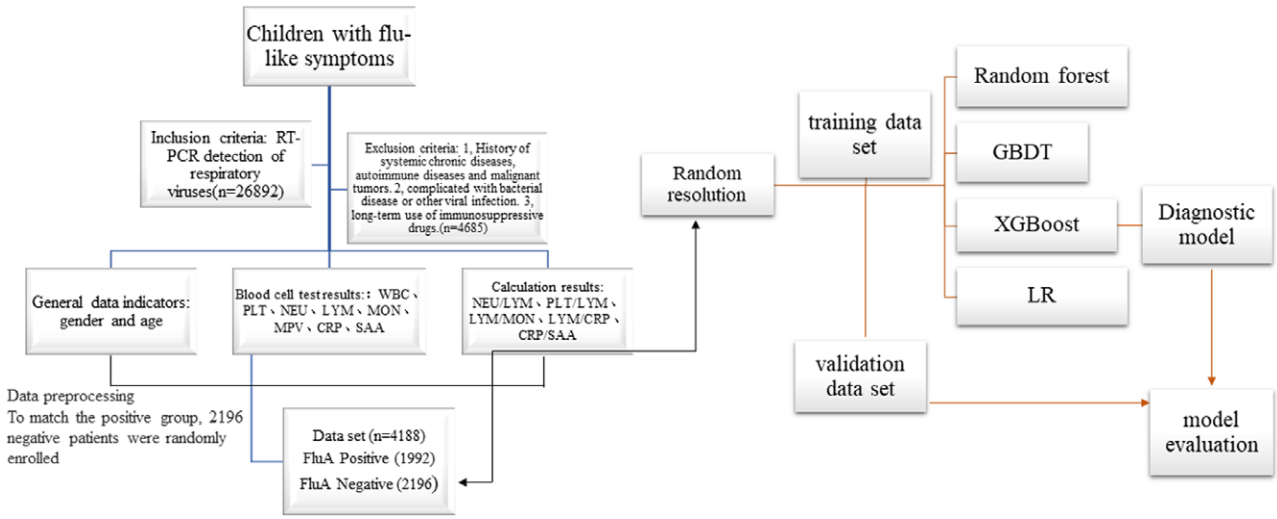

Figure 1. Flow diagram.

Table 2

Evaluation indexes of various algorithms.

	AUC (95%CI)	Sensitivity (%)	Specificity (%)	NPV (%)	PPV (%)
Random forest algorithm	0.835 (0.815–0.854)	75.10	75.13	79.12	70.63
GBDT	0.872 (0.856–0.888)	77.23	80.29	80.47	77.03
XGBoost	0.867 (0.850–0.884)	76.28	79.66	78.58	77.44
Logistic regression	0.784 (0.762–0.805)	67.28	74.53	70.59	71.48

AUC = area under the curve, CI = confidence interval, NPV = negative predictive value, PPV = positive predictive value.

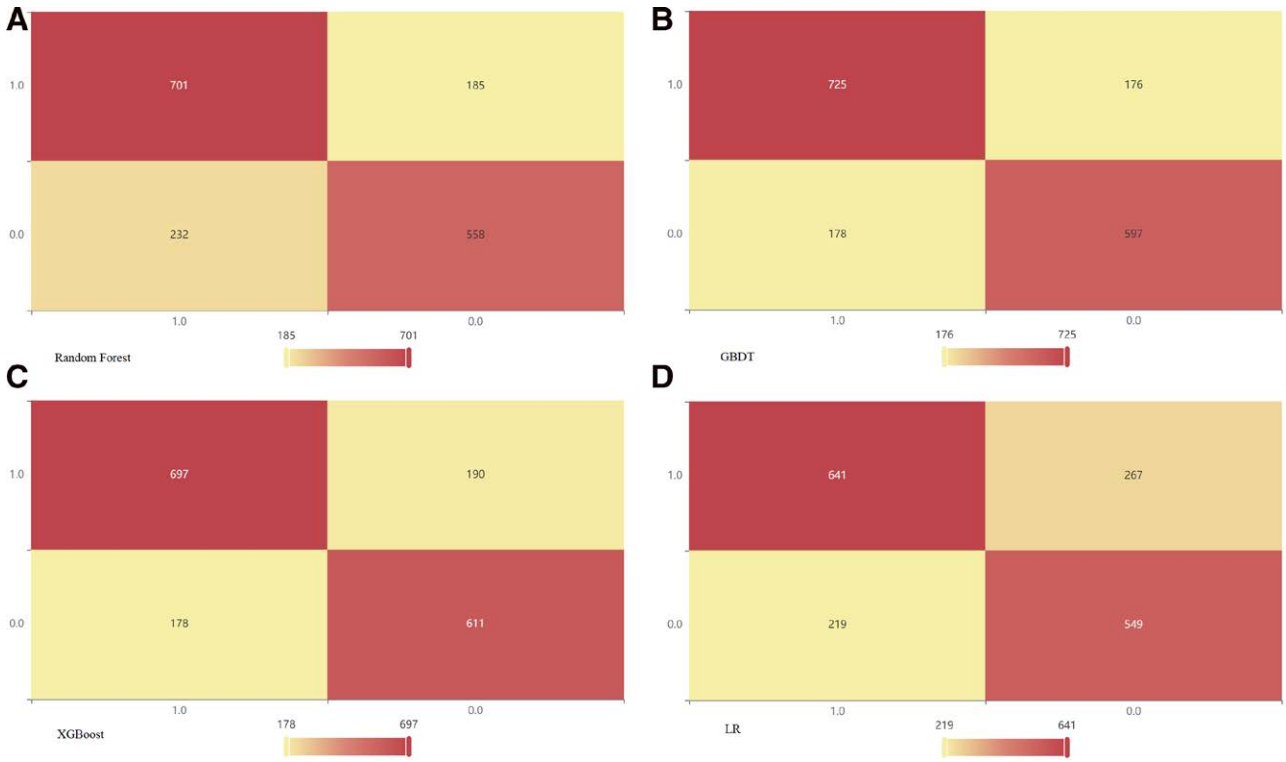

Figure 2. Heatmap of test data.

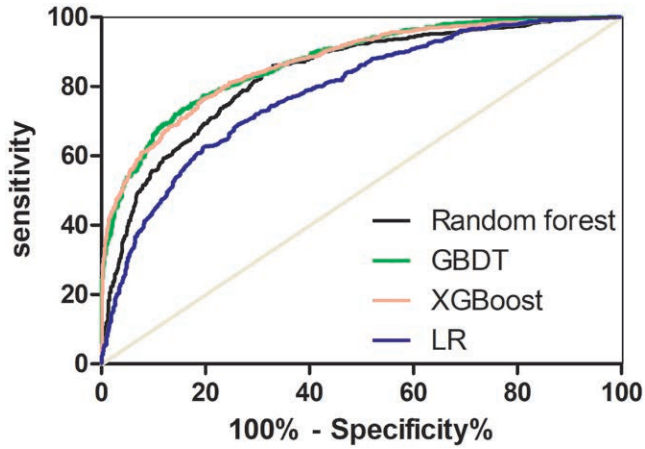

Figure 3. ROC curves of algorithm.

others. Zimmerman et al found a sensitivity of 84%, specificity of 49%, and AUROC of 0.69 in their prediction model.[24] Anderson et al achieve a sensitivity of 52.1%, specificity of 82.9%, and AUROC of 0.689.[25] Therefore, our model can effectively distinguish between children with influenza A and influenza like symptoms, and the diagnostic level is close to a clinical physician. LR as a traditional machine learning algorithm, its diagnostic ability is slightly worse than the other 3 models, AUC is 0.784. From the perspective of diagnostic accuracy, building diagnostic models on existing data is feasible and has high accuracy and performance.

Combined with the 4 modeling algorithms, it can be concluded that variables such as LYM count, age, SAA, WBC count and PLR may play an important role in predicting influenza A. They showed significant differences ($p<0.001$) in distinguishing between influenza A and influenza-like illness, which was similar to previous studies.[26,27] LYM and WBC counts were significantly reduced in children with influenza A, while age, SAA, and PLR were significantly increased compared to children with influenza-like diseases. This provides a preliminary diagnosis and treatment direction for clinical doctors. When influenza virus invades the body lymphocytes specifically bind to influenza virus and transfer to inflammatory sites, resulting in apoptosis of peripheral blood lymphocytes and a decrease in the total number. Due to the effect of influenza virus in human body, lymphocytopenia, as well as the complex interaction between platelets and immune cells in vivo, their extensive involvement in inflammatory reactions and the adhesion and aggregation of neutrophil and pulmonary microvascular endothelial cells lead to the reduction of WBC count.[28–30]

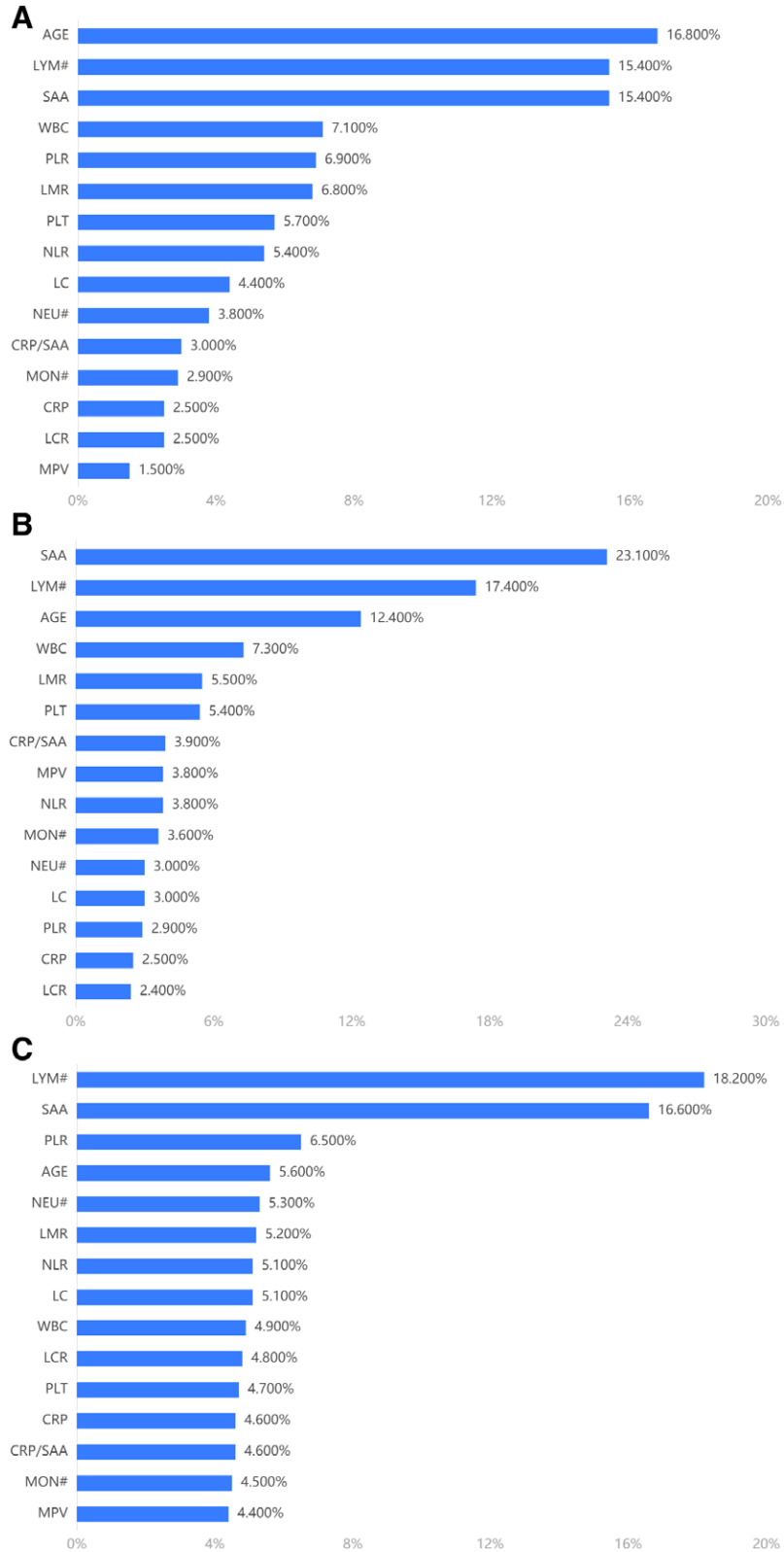

Figure 4. Feature importance of the algorithm.

This model only requires basic information of the patient and blood routine results, which can be quickly obtained at all levels of hospitals. Compared to RT-PCR, it is simpler and faster, which is beneficial for all medical institutions to obtain accurate influenza A diagnosis information with the simplest results, especially suitable for grassroots medical units. The effective utilization of the model can help doctors determine whether patients should have nasopharyngeal swab test, help to carry out pre-diagnosis of infectious diseases, shortens the treatment process, and reduces economic burden. Given the good

diagnostic ability of the model, we will launch a mini program on the hospital system for trial, collect clinical trial results and feedback information, and conduct external validation.

However, there are some limitations in our research: There are missing values in the case data, although we have supplemented the data with an average, it is still possible to lead to the result bias. Due to the limited amount of data included, more data should be collected to further optimize the diagnostic model for influenza A in children.

In conclusion, we relied on computer algorithms to establish a model that can quickly, accurately, and economically diagnose influenza A in children. This model has good transferability and wide applicability, making it very suitable for promotion in grassroots hospitals. The successful establishment of the model can early diagnose children with influenza, shorten diagnosis and treatment time, reduce the incidence of complications and severe cases, and reduce the economic burden for billions of families.

Author contributions

Conceptualization: Qian Zeng.
Data curation: Qian Zeng, Chun Yang, Yurong Li.
Methodology: Xin Lv.
Writing – original draft: Qian Zeng.
Writing – review & editing: Xinran Geng, Xin Lv.

References

[1] Nayak J, Hoy G, Gordon A. Influenza in children. Cold Spring Harbor Perspect Med. 2021;11.
[2] Piroth L, Cottenet J, Mariet AS, et al. Comparison of the characteristics, morbidity, and mortality of COVID-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. Lancet Respiratory Med. 2021;9:251–9.
[3] Xu C, Liu L, Ren B, et al. Incidence of influenza virus infections confirmed by serology in children and adult in a suburb community, northern China, 2018-2019 influenza season. Influenza Other Respiratory Viruses. 2021;15:262–9.
[4] Iuliano AD, Roguski KM, Chang HH, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet. 2018;391:1285–300.
[5] Leung CH, Tseng HK, Wang WS, et al. Clinical characteristics of children and adults hospitalized for influenza virus infection. J Microbiol Immunol Inf. 2014;47:518–25.
[6] Lafond KE, Nair H, Rasooly MH, et al. Global role and burden of influenza in pediatric respiratory hospitalizations, 1982-2012: a systematic analysis. PLoS Med. 2016;13:e1001977.
[7] Zheng J, Huo X, Huai Y, et al. Epidemiology, seasonality and treatment of hospitalized adults and adolescents with influenza in Jingzhou, China, 2010-2012. PLoS One. 2016;11:e0150713.
[8] Cho CH, Woo MK, Kim JY, et al. Evaluation of five rapid diagnostic kits for influenza A/B virus. J Virol Methods. 2013;187:51–6.
[9] Arai J, Aoki T, Sato M, et al. Machine learning-based personalized prediction of gastric cancer incidence using the endoscopic and histologic findings at the initial endoscopy. Gastrointest Endosc. 2022;95:864–72.
[10] Liu R, Wang M, Zheng T, et al. An artificial intelligence-based risk prediction model of myocardial infarction. BMC Bioinformatics. 2022;23:1–17.
[11] Liao Y, Liu C, He W, et al. Study on the value of blood biomarkers NLR and PLR in the clinical diagnosis of influenza a virus infection in children. Clin Lab. 2021;67:2540–7.
[12] Zhu R, Chen C, Wang Q, et al. Routine blood parameters are helpful for early identification of influenza infection in children. BMC Infect Dis. 2020;20:864.
[13] Coskun O, Avci IY, Sener K, et al. Relative lymphopenia and monocytosis may be considered as a surrogate marker of pandemic influenza a (H1N1). J Clin Virol. 2010;47:388–9.
[14] Fei Y, Zhang H, Zhang C. The application of lymphocyte*platelet and mean platelet volume/platelet ratio in influenza A infection in children. J Clin Lab Anal. 2019;33:e22995.
[15] Vasileva D, Badawi A. C-reactive protein as a biomarker of severe H1N1 influenza. Inflamm Res. 2019;68:39–46.
[16] Zimmerman O, Rogowski O, Aviram G, et al. C-reactive protein serum levels as an early predictor of outcome in patients with pandemic H1N1 influenza A virus infection. BMC Infect Dis. 2010;10:288.
[17] Yao Z, Zhang Y, Wu H. Regulation of C-reactive protein conformation in inflammation. Inflammation Res. 2019;68:815–23.
[18] Gao R, Wang L, Bai T, et al. C-Reactive protein mediating immunopathological lesions: a potential treatment option for severe influenza A diseases. EBioMedicine. 2017;22:133–42.
[19] Mo X N SZQ, Lei CL. Serum amyloid A is a predictor for prognosis of COVID-19 %. J Respirol. 2020;25:764–5.
[20] Kumar V. Influenza in children. Indian J Pediatr. 2017;84:139–43.
[21] Ozsurekci Y, Aykac K, Bal F, et al. Outcome predictors of influenza for hospitalization and mortality in children. J Med Virol. 2021;93:6148–54.
[22] Ocal Demir S, Atici S, Kepenekli Kadayifci E, et al. Influenza A (H1N1)-associated severe complications; hemolytic uremic syndrome, myocarditis, acute necrotizing encephalopathy. J Infect Dev Countries. 2019;13:83–6.
[23] ; Committee on Infectious D. Recommendations for prevention and control of influenza in children, 2020-2021. Pediatrics. 2020;146.
[24] Zimmerman RK, Balasubramani GK, Nowalk MP, et al. Classification and Regression Tree (CART) analysis to predict influenza in primary care patients. BMC Infectious Diseases. 2016;16:1–11.
[25] Anderson KB, Sriluck S, Veerachai W, et al. Clinical and laboratory predictors of influenza infection among individuals with influenza-like illness presenting to an urban Thai hospital over a five-year period. PLoS One. 2018;13:e0193050.
[26] Patel B, Oye M, Norez D, et al. Peripheral blood lymphocyte-to-monocyte ratio as a screening marker for influenza infection. J Investig Med. 2021;69:47–51.
[27] Temel H, Gunduz M, Tosun AI, et al. The importance of Neutrophil/Lymphocyte and Lymphocyte/Monocyte ratios in the diagnosis of influenza in children. Clin Lab. 2021;67:1073–8.
[28] Wang L, Chang LS, Lee IK, et al. Clinical diagnosis of pandemic A(H1N1) 2009 influenza in children with negative rapid influenza diagnostic test by lymphopenia and lower C-reactive protein levels. Influenza Other Respiratory Viruses. 2014;8:91–8.
[29] Sugiyama MG, Gamage A, Zyla R, et al. Influenza virus infection induces platelet-endothelial adhesion which contributes to lung injury. J Virol. 2016;90:1812–23.
[30] Rudd JM, Pulavendran S, Ashar HK, et al. Neutrophils induce a novel chemokine receptors repertoire during influenza pneumonia. Front Cell Infect Microbiol. 2019;9:108.