


RESEARCH

Open Access



Asthma prediction via affinity graph enhanced classifier: a machine learning approach based on routine blood biomarkers

Dejing Li^{1†}, Stanley Ebhohimhen Abhadiomhen^{2,3†}, Dongmei Zhou⁴, Xiang-Jun Shen², Lei Shi^{5*} and Yubao Cui^{4*} 

Abstract

Background Asthma is a chronic respiratory disease affecting millions of people worldwide, but early detection can be challenging due to the time-consuming nature of the traditional technique. Machine learning has shown great potential in the prompt prediction of asthma. However, because of the inherent complexity of asthma-related patterns, current models often fail to capture the correlation between data samples, limiting their accuracy. Our objective was to use our novel model to address the above problem via an Affinity Graph Enhanced Classifier (AGEC) to improve predictive accuracy.

Methods The clinical dataset used in this study consisted of 152 samples, where 24 routine blood markers were extracted as features to participate in the classification due to their ease of sourcing and relevance to asthma. Specifically, our model begins by constructing a projection matrix to reduce the dimensionality of the feature space while preserving the most discriminative features. Simultaneously, an affinity graph is learned through the resulting subspace to capture the internal relationship between samples better. Leveraging domain knowledge from the affinity graph, a new classifier (AGEC) is introduced for asthma prediction. AGECE's performance was compared with five state-of-the-art predictive models.

Results Experimental findings reveal the superior predictive capabilities of AGECE in asthma prediction. AGECE achieved an accuracy of 72.50%, surpassing FWAdaBoost (61.02%), MLFE (60.98%), SVR (64.01%), SVM (69.80%) and ERM (68.40%). These results provide evidence that capturing the correlation between samples can enhance the accuracy of asthma prediction. Moreover, the obtained *p* values also suggest that the differences between our model and other models are statistically significant, and the effect of our model does not exist by chance.

Conclusion As observed from the experimental results, advanced statistical machine learning approaches such as AGECE can enable accurate diagnosis of asthma. This finding holds promising implications for improving asthma management.

Keywords Asthma prediction, Asthma, Affinity graph, Feature selection

[†]Dejing Li and Stanley Ebhohimhen Abhadiomhen contributed equally to the paper.

*Correspondence:

Lei Shi

13818226306@139.com

Yubao Cui

ybcui1975@hotmail.com

Full list of author information is available at the end of the article



Background

Asthma affects 235 million people globally [1], making it one of the most common chronic diseases in the world, according to the World Health Organization [2]. Specifically, asthma is characterized by inflammation of the airways, which results in symptoms such as wheezing, shortness of breath, and chest tightness [3, 4]. In order to avoid exacerbations and hospitalizations, asthma must be accurately and promptly diagnosed for effective management and treatment of the disease [5]. Conventional diagnostic methods often combine medical history, physical examination, and lung function tests. Apart from the fact that these tests are expensive, atypical symptoms in some patients can result in delayed or missed diagnoses. Moreover, asthma in young children can be very difficult to diagnose, and traditional methods may exacerbate the situation due to their time-consuming nature [6].

With the advancement of machine learning (ML), there is a growing interest [7–13] in predicting asthma using computational techniques to analyze medical data, identify patterns and generate predictions that can assist healthcare providers in early and more accurate diagnoses of asthma. Typical predictive models include Decision Trees [14], Random Forests [15], Support Vector Machines (SVMs) [16], Neural Networks [17], and Bayesian Networks [18]. Despite the successes of these classical ML models, they often cannot capture the internal relationships between data samples, making them less robust for complex medical conditions like asthma. This inadequacy could arise from a combination of limitations in model complexity, algorithmic constraints, and insufficient adaptability to dynamic and intricate patterns within the asthma data. Addressing this problem may help unlock the full potential of ML in the prediction and management of asthma. Recently, graph-based learning (GBL) [19, 20] has emerged as a promising method for capturing correlation between data samples. GBL has found widespread use in subspace clustering [21–23] via an affinity graph construction. Here, each sample is reconstructed by a linear combination of other samples in the same subspace. According to Lu et al. [24], such subspace representation can allow for a more detailed understanding of data and can reveal important patterns that might be missed by traditional clustering methods.

Inspired by this, a new ML approach, which uses an affinity graph enhanced classifier (AGEC) for asthma

prediction, is proposed in this paper. As far as we know, this is the first study that directly exploits an affinity graph for classification. Accordingly, we demonstrate through experimental evaluation with existing ML models that AGECE can tackle the above problem and improve asthma prediction accuracy. Therefore, we hope that the results of our study can assist the clinical community in the prompt prediction and management of asthma.

Methods

Data collection

The datasets used in this study contained 152 records collected from asthma patients in the Affiliated Shuguang Hospital of Shanghai Traditional Chinese Medicine University. Before the study was conducted, ethical approval was obtained from the relevant ethics committee at the Affiliated Shuguang Hospital of Shanghai Traditional Chinese Medicine University. The sample population in the dataset ranges between 20 and 100 years old. Of the 152 samples in the dataset, 18.4% are between 20 and 40 years old, 47.4% are between 50 and 69 years old, and 34.2% are over 70 years old. The age distribution of the sample indicates that the majority of the participants were between 50 and 69 years old. In terms of gender, the dataset includes 40% males and 60% females, with a male to female ratio of roughly 4:6 (see Table 1 for a summary of the dataset). For each record, twenty-four indicators which include complete blood count differentials and red blood cell indices were extracted for use as candidate predictors in the classification procedure, as shown in Table 2. The diagnosis results were used as the label. In this study, there are five possible diagnosis categories: asthma, bronchial asthma, sputum turbidiosis, non-critical-bronchial asthma, and no diagnosis.

Model formulation

This section describes the formulation of our proposed model. Firstly, in order to transform the raw data into appropriate format that can be used by the model, we represented the input dataset $X = [x_1, x_2, x_3, \dots, x_n] \in \mathbb{R}^{p \times n}$, and the label set $Y \in \{1, 0\}^{q \times n}$, where q denotes the label dimension, p denotes the feature dimension, and n represents the number of samples. For such representation, the traditional multi-label learning [25] adopts the binary linear regression model to learn matrix $W_{p \times q}$, as follows:

Table 1 Summary of the characteristics of the dataset

Samples	Features	% of men	% of woman	Age distribution of samples	% of samples with age between 20–40 years	% of samples with age between 50–69 years	% of samples with age above 70 years
152	24	40	60	Between 20–100 years	18.4	47.4	34.2

Table 2 Twenty-four clinical indicators extracted as candidate predictors

White blood cell (WBC)	Neutrophil percentage (NE%)	Lymphocyte percentage (LY%)	Monocyte percentage (MO%)	Eosinophil percentage (EO%)	Basophil percentage (BA%)	Neutrophil count—absolute (NE#)	Lymphocyte count—absolute (LY#)
Monocyte count—absolute (MO#)	Eosinophil count—absolute (EO#)	Basophil count—absolute (BA#)	Red blood cell (RBC)	Hemoglobin (HGB)	Hematocrit (HCT)	Mean corpuscular volume (MCV)	Mean corpuscular hemoglobin (MCH)
Mean corpuscular hemoglobin concentration (MCHC)	Red cell distribution width (RDW)	Platelet count (PLT)	Platelet distribution width (PDW)	Platelet crit (PCT)	Mean platelet volume (MPV)	C-reactive protein (CRP)	Serum amyloid A (SAA)

$$\arg \min \left\| Y - W^T X \right\|_F^2 \tag{1}$$

However, the model has many shortcomings. When the label dimension is large, its accuracy will be reduced. At the same time, the model ignores the correlation between samples. Aiming at this problem, a new model was constructed in this study. To aid easy understanding, the model formulation is divided into several steps as follows.

Capturing the correlation between samples

To capture internal relation between samples and improve the classification effect of the traditional multi-label model in asthma prediction, we considered using domain information from the sample to enhance robustness. To arrive at our model, a projection matrix P was obtained first to reduce the dimensionality of the feature space and preserve the most discriminative features so that similar sample nodes are closer to each other and their corresponding label nodes are also close to each other. Simultaneously, an affinity graph W was learned on the resulting subspace to capture the domain information. The specific formula is as follows:

$$\sum_{i,j} \left\| p^T x_i - p^T x_j \right\|_F^2 W_{ij} \Leftrightarrow 2tr \left(P^T X L X^T P \right) \tag{2}$$

where the projection matrix is obtained, such that $P^T X \rightarrow Y$. In order to avoid trivial solutions, we imposed nonnegative and normalized constraints on the graph. Therefore, the above model was transformed into:

$$\sum_{i,j} \left\| p^T x_i - p^T x_j \right\|_F^2 W_{ij} + \|W\|_F^2 \tag{3}$$

$s.t. 0 \leq W_{ij} \leq 1, \sum_j W_{ij} = 1$

Specifically, by introducing the affinity matrix W , we can further learn the relationship between samples. The

value of the W matrix represents the degree of correlation between the similar sample and samples from other classes. That is, the closer the distance between sample nodes, the greater the correlation.

Affinity graph enhanced classifier

As depicted by Eq. (3), P projects the original feature space into the low-dimensional space to reduce the number of digits in the feature space. The affinity graph is learned on the low-rank subspace to capture the correlation between samples. On this basis, a new classifier Z was constructed to benefit from the domain information through the affinity graph. This strategy helps uncover complex data patterns that hold clinical relevance in the context of asthma. In addition, in order to avoid redundant information in the feature space and make the low-dimensional mapping of data retain the main information in the original data, we introduced an orthogonal constraint $P^T X X^T P = I$, and the new optimization model became:

$$\min \|Y - ZW\|_F^2 + \lambda_2 \|Z\|_F^2 \quad s.t. P^T X X^T P = I \tag{4}$$

Furthermore, we introduced an auxiliary variable M through the constraint $W = M$ to make Eq. (4) easier to solve, similar to the previous works [26, 27]. Therefore, combining Eqs. (3) and (4), our objective function was obtained as:

$$\min \|Y - ZM\|_F^2 + \lambda_1 \sum_{i,j} \left\| p^T x_i - p^T x_j \right\|_F^2 W_{ij} + \lambda_2 \|Z\|_F^2 + \lambda_3 \|W\|_F^2 \tag{5}$$

$s.t. 0 \leq W_{ij} \leq 1, \sum_j W_{ij} = 1, W = M, P^T X X^T P = I$

where, λ_1 , λ_2 and λ_3 denote the regularization parameters used to constrain the second, third, and fourth terms. Figure 1 describes the framework of the proposed method.

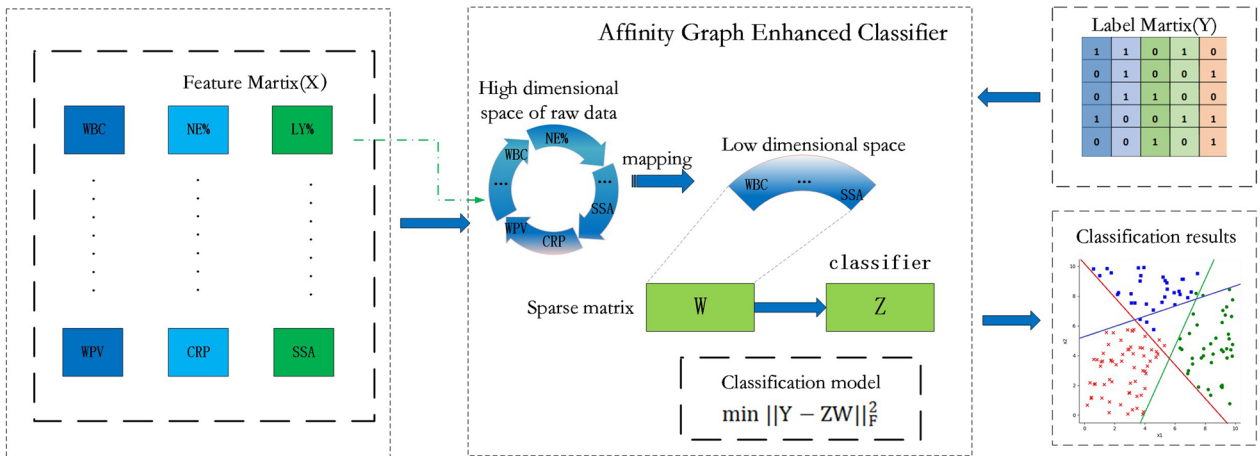


Fig. 1 Framework of the proposed method. As can be seen in the figure, the original data is mapped first into a low-dimensional space. A classifier is then constructed to leverage the domain information from the affinity graph for asthma prediction

Model optimization

In order to solve our objective function, an efficient optimization algorithm was implemented based on the Augmented LaGrange Multiplier (ALM) strategy [28]. Before that, we obtained the Augmented LaGrange function as follows.

$$\min \|Y - ZM\|_F^2 + \lambda_1 \sum_{i,j} \left| |p^T x_i - p^T x_j| \right|_F^2 W_{ij} + \lambda_2 \|Z\|_F^2 + \lambda_3 \|W\|_F^2 + tr(Y_1^T (W - M)) + \frac{\mu_1}{2} \|W - M\|_F^2 \tag{6}$$

where Y_1 is the LaGrange multiplier, which is necessary for solving constrained problems. Thus, separating the unconnected terms in Eq. (6), the minimization problem and the ideal solution for each variable are given below in no particular order.

Z subproblem

Considering only the terms containing Z, we obtained the following optimization function.

$$\min \|Y - ZM\|_F^2 + \lambda_2 \|Z\|_F^2 \tag{7}$$

Thus, expanding the first item in Eq. (7), we arrived at:

$$\|Y - ZM\|_F^2 = tr((Y - ZM)^T (Y - ZM)) = tr(Y^T Y - Y^T ZM - M^T Z^T Y - M^T Z^T ZM) \tag{8}$$

After considering only variable Z, we obtained:

$$\min tr(M^T Z^T ZM - 2M^T Z^T Y) + \lambda_2 tr(Z^T Z) \tag{9}$$

Consequently, a partial derivative of Z yielded:

$$\partial_Z = (MM^T Z^T - MY^T + \lambda_2 Z^T)^T = \lambda_2 Z + ZMM^T - YM^T \tag{10}$$

Setting the Eq. (10) equal to 0, that is, $\lambda_2 Z + ZMM^T - YM^T = 0$, the optimal solution of Z was obtained through the following formula:

$$Z = YM^T (\lambda_2 I + MM^T)^{-1} \tag{11}$$

P subproblem

$$\min \lambda_1 \sum_{i,j} \left| |p^T x_i - p^T x_j| \right|_F^2 W_{ij} \quad \text{s.t. } P^T X X^T P = I \tag{12}$$

Expanding the above optimization function, Eq. (12) can be rewritten as:

$$\min \lambda_1 tr(P^T X L_b X^T P) \quad \text{s.t. } P^T X X^T P = I \tag{13}$$

Therefore, using Lagrange multiplier method, we obtained:

$$\min \lambda_1 tr(P^T X L_b X^T P) - (P^T X X^T P - I) \tag{14}$$

A partial derivative of P yielded:

$$\partial_P = (\lambda_1 P^T X L_b X^T - (P^T X X^T))^T \tag{15}$$

Setting Eq. (15) equal to 0,

$$\lambda_1 P = (X L_b X^T)^{-1} X X^T P \tag{16}$$

Finally, the optimal value of matrix P was obtained by finding the eigenvector corresponding to matrix $(XL_bX^T)^{-1}XX^T$.

M subproblem

$$\min \|Y - ZM\|_F^2 + \text{tr}(Y_1^T(W - M)) + \frac{\mu_1}{2}\|W - M\|_F^2 \quad (17)$$

As mentioned previously, Y_1 is the Lagrange multiplier, and $\mu_1 > 0$ is the penalty parameter. Equation (17) can be rewritten as:

$$\begin{aligned} \min & \text{tr}((Y - ZM)^T(Y - ZM)) + \text{tr}(Y_1^T(W - M)) + \frac{\mu_1}{2}\|W - M\|_F^2 \\ & = \min \text{tr}(Y^TY - Y^TZM - M^TZ^TY - M^TZ^TZM) + \text{tr}(Y_1^T(W - M)) \\ & \quad + \frac{\mu_1}{2}((W - M)^T(W - M)) \end{aligned} \quad (18)$$

Extracting only variables related to M :

$$\min \text{tr}(M^TZ^TZM - 2M^TZ^TY) + \text{tr}(Y_1^TM) + \frac{\mu_1}{2}(M^TM - 2W^TM) \quad (19)$$

As with the other variables, a partial derivative of M yielded:

$$\partial_M = (2M^TZ^TZ - 2Y^TZ - Y_1^T + \mu_1M^T - \mu_1W^T)^T \quad (20)$$

Setting Eq. (20)=0, the optimal solution of M was obtained through the following formula:

$$M = \left(\frac{2}{\mu_1}Z^TZ + I\right)^{-1} \left(W + \frac{1}{\mu_1}Y_1 + \frac{2}{\mu_1}Z^TY\right) \quad (21)$$

W subproblem

$$\begin{aligned} \sum_{ij} \left\| p^T x_i - p^T x_j \right\|_F^2 W_{ij} + \lambda_3 \|W\|_F^2 \\ + \text{tr}(Y_1^T(W - M)) + \frac{\mu_1}{2}\|W - M\|_F^2 \end{aligned} \quad (22)$$

Expanding Eq. (22), we arrived at:

$$\sum_{ij} \left\| p^T x_i - p^T x_j \right\|_F^2 W_{ij} + \frac{\mu_1}{2} \left\| W - M + \frac{Y_1}{\mu_1} \right\|_F^2 + \lambda_3 \|W\|_F^2 \quad (23)$$

By making $M - \frac{Y_1}{\mu_1} = U$, Eq. (23) can be rewritten as:

$$\begin{aligned} \min_{W_i} \sum_{ij} \left\| p^T x_i - p^T x_j \right\|_F^2 W_{ij} + \frac{\mu_1}{2} \|W - U\|^2 + \lambda_3 \|W\|_F^2 \\ \text{s.t. } 0 \leq W_{ij} \leq 1, \sum_j W_{ij} = 1 \end{aligned} \quad (24)$$

Because Eq. (24) is independent for each i , we solved W_i separately as follows:

$$\begin{aligned} \min_{W_i} \sum_{ij} \left\| p^T X_i - p^T X_j \right\|^2 W_{ij} + \frac{\mu_1}{2} \|W_i - U_i\|^2 + \lambda_3 \|W_i\|^2 \\ \text{s.t. } 0 \leq W_{ij} \leq 1, \sum_j W_{ij} = 1 \end{aligned} \quad (25)$$

Denoting $e_{ij} = \|p^T X_i - p^T X_j\|$, $w_v = \frac{\mu_1}{2}$, we rewrite Eq. (25) in the following way.

$$\begin{aligned} \min_{W_i} \frac{1}{2} \left\| W_i + \frac{e_i}{2\lambda_3} \right\|^2 + \frac{1}{2\lambda_3} \|U_i - W_i\|^2 \\ \text{s.t. } 0 \leq W_{ij} \leq 1, \sum_j W_{ij} = 1 \end{aligned} \quad (26)$$

$$\begin{aligned} L(W_i, \eta, \xi) = \frac{1}{2} \left\| W_i + \frac{e_i}{2\lambda_3} \right\|^2 + \frac{1}{2\lambda_3} \|U_i - W_i\|^2 \\ - \eta(1^T W_i - 1) - \xi^T W_i \end{aligned} \quad (27)$$

η is the scalar of the Lagrange coefficient, and ξ is the vector of the Lagrange coefficient. Taking a partial derivative of W_i , we obtained:

$$W_i + \frac{e_i}{2\lambda_3} - \frac{1}{\lambda_3} w_v (U_i - W_i) - \eta 1 - \xi = 0 \quad (28)$$

The j th term of W_i in the equation is:

$$W_{i,j} + \frac{e_{i,j}}{2\lambda_3} - \frac{1}{\lambda_3} w_v (U_{i,j} - W_{i,j}) - \eta - \xi_j = 0 \quad (29)$$

By following the KKT conditions [29], we obtained $W_{i,j}$ through the following formula.

$$W_{i,j} = \left(\frac{-\frac{e_{i,j}}{2} + w_v U_{i,j} + \lambda_4 \eta}{\lambda_3 + w_v} \right) + \quad (30)$$

Furthermore,

$$-\frac{e_{i,k}}{2} + w_v U_{i,k} + \lambda_3 \eta > 0 \text{ and } -\frac{e_{i,k+1}}{2} + w_v U_{i,k+1} + \lambda_3 \eta \leq 0, \eta = \frac{1}{k} \left(1 + \frac{w_v}{\lambda_3} + \sum_{h=1}^k \frac{e_{i,h}}{2\lambda_3} \right),$$

$$\begin{cases} \lambda_4 > \frac{ke_{i,k} - \sum_{h=1}^k e_{i,h} - 2kw_v U_{i,k} - 2w_v}{2} \\ \lambda_4 \leq \frac{ke_{i,k+1} - \sum_{h=1}^k e_{i,h} - 2kw_v U_{i,k+1} - 2w_v}{2} \end{cases}$$

$$\lambda_3 = \frac{ke_{i,k+1} - \sum_{h=1}^k e_{i,h} - 2kw_v U_{i,k+1} - 2w_v}{2}.$$

$$W_{i,j} = \begin{cases} \frac{e_{i,k+1} - e_{i,j} + 2w_v U_{i,j} - 2w_v U_{i,k+1}}{ke_{i,k+1} - \sum_{h=1}^k e_{i,h} - 2kw_v U_{i,k+1} + 2\sum_{h=1}^k w_v U_{i,h}}, & j \leq k \\ 0, & j > k \end{cases} \quad (31)$$

For the detailed derivation and proof of Eq. (31), refer to reference [30]. A summary of the complete solution of our proposed model is captured in Algorithm1.

Compared classification algorithms

Five classification algorithms were used to build classification models for comparison with our AGECE model. They are, multi-label learning with feature-induced labeling information enrichment (MLFE) [31], support vector machines (SVM), exclusivity regularized machine (ERM) [32], support vector regression (SVR) [33], and multi-class fuzzily weighted AdaBoost (FWAdaBoost) [34]. We considered these algorithms for comparison because they use a similar strategy to AGECE or because they are often used for building asthma predictive models. For example, MLFE is a multi-label learning algorithm like ours. SVM and SVR are commonly used for building asthma predictive models due to their excellent generalization ability [35]. ERM and FWAdaBoost, which is based on AdaBoost [36] uses the ensemble learning strategy, which is well-known to improve the performance of single-task learning models.

Algorithm 1 The Algorithm of the proposed model

Input: Data \mathbf{X}, Y , parameter $\lambda_1, \lambda_2, \lambda_3$

Initialize $P = Z = W = 0, Y_1 = 0, \rho = 1.1; \mu = 10^{-4}; \mu_{max} = 10^6; \epsilon = 10^{-4};$

While not converged **do**

- 1 Fix others and update Z by Eq. (11);
- 2 Fix others and update P by Eq. (16);
- 3 Fix others and update M by Eq. (21);
- 4 Fix others and update W by Eq. (30);
- 5 Update the multiplier.

$$Y_1 = Y_1 + \mu(M - W)$$

- 5 Update μ by $\mu = \min(\rho\mu, \max(\mu))$
- 6 Check the convergence conditions:

$$\|M - W\|_{\infty} < \epsilon$$

Output: Z, P, W

Table 3 ACC of AGE C compared with different models

	MLFE	SVM	ERM	SVR	FWAdaBoost	AGE C
ACC	0.6098	0.6980	0.6840	0.6401	0.6102	0.7250

The value in bold font symbolizes the best performance

Table 4 AUC and P values of AGE C compared with different models

Algorithm	Prediction accuracy (area under the curve)	
	AUC	P value
MLFE	0.5201 ± 0.016	0.0304
SVM	0.7034 ± 0.012	0.0302
ERM	0.6632 ± 0.030	0.0298
SVR	0.6312 ± 0.022	0.0305
FWAdaBoost	0.7014 ± 0.002	0.0301
AGE C	0.7401 ± 0.021	0.0305

The value in bold font symbolizes the best performance

Evaluation

The experimental results were captured in terms of accuracy (ACC) and the area under the receiver operating characteristic (ROC) curve (AUC). These metrics were utilized to characterize and compare the performance of

the various classification algorithms in asthma prediction. While ACC measures how well a model can correctly predict class labels of the instances in the test set, AUC measures the overall performance of a classifier by evaluating its ability to distinguish between positive and negative instances. Unlike ACC, AUC is insensitive to changes in class distribution.

Results

Experiment settings

The comparison algorithms and our AGE C algorithm were implemented using MATLAB R2016a installed on a Windows 10 computer system. In order to reasonably evaluate the effectiveness of our model, two sets of experiments were performed. The first set examined the performance of each algorithm using all 24 clinical indicators. The second investigated the effect of different subsets of the features on the performance of the proposed method. In each experiment, we first divided the dataset into a training set and a held-out testing

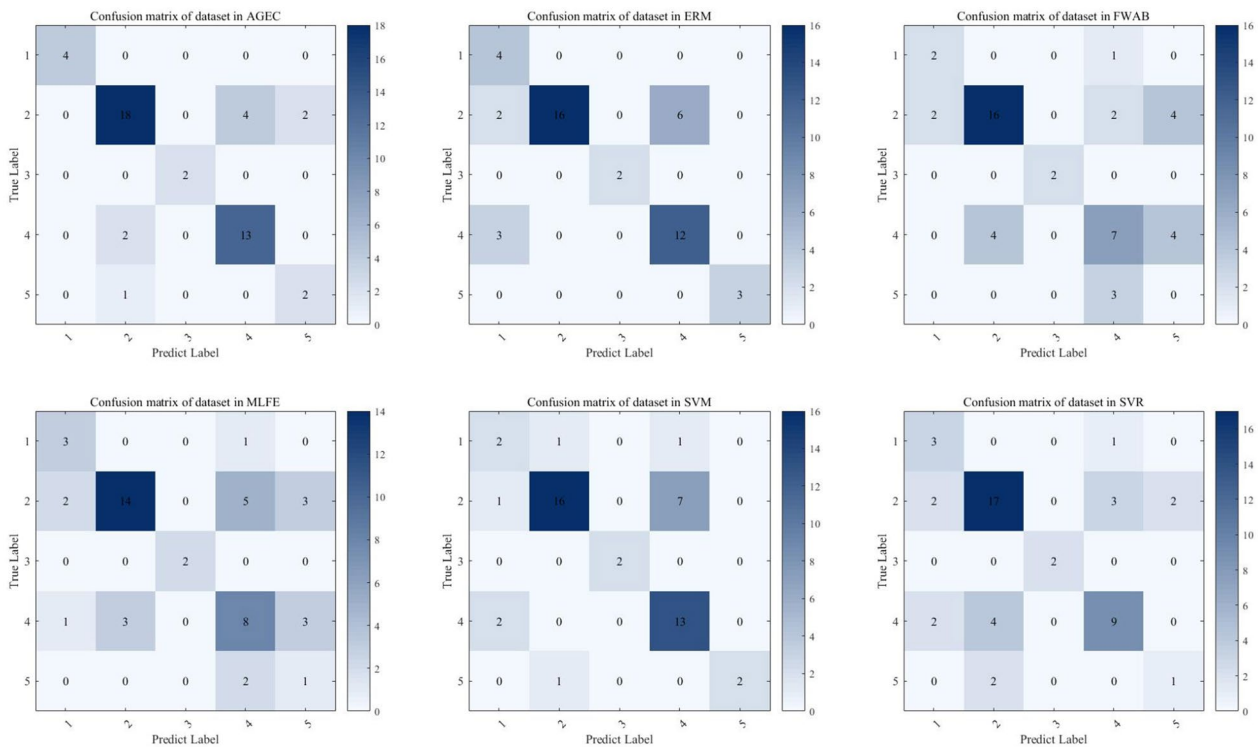


Fig. 2 The confusion matrix obtained for each of the six approaches

set with a ratio of 1:1. Then twofold cross-validation was performed on the training dataset for parameter tuning. We selected 2 based on the relatively small size of our dataset. Moreover, the grid search strategy was also applied to tune the hyperparameters during cross-validation. The optimal hyperparameters for our best AGECE were $\lambda_1 = 8 * 10^{-4}$, $\lambda_2 = 2 * 10^{-5}$, $\lambda_3 = 1.8 * 10^{-4}$.

Evaluation of the prediction models

Table 3 displays the performance in terms of the accuracy of various models, including AGECE, in asthma prediction. As can be seen from the results, AGECE obtained an accuracy of 72.50%, which is significantly higher than other models. Although there is a seemingly insignificant gap of 2.7% between AGECE and the SVM model, the gap widens in terms of AUC, as shown in Table 4. Specifically, AGECE obtained an AUC of 74.01%, which is significantly higher than SVM by over 3% and much higher than the other models. This suggests that our model has the better capability in distinguishing between asthmatic and non-asthmatic patients. Moreover, the *p* value also suggest that the differences between our model and other models are statistically significant, and the effect of our model does not exist by chance. In addition, to more specifically demonstrate the advantages of our proposed model, Fig. 2 shows the confusion matrix obtained for each of the six models. As can be seen in the figure, the shadow on the diagonal of our AGECE is deeper than that on other models, which means that our model can make more correct classification results than other models. Meanwhile, the shadow on the non-diagonal is less than that on other models, which means that our model can predict fewer wrong results.

Additionally, we also conducted comparison with some regression models: Logistic Regression, Random Forest (RF) and Lasso. The results, as presented in Table 5, indicate that the accuracy of Logistic Regression (59.24%), RF (54.21%), and Lasso (56.01%) is notably lower than the accuracy achieved by the previously compared methods. This comparison highlights the superior performance of our proposed method in the context of asthma prediction. Moreover, the observed lower accuracy of Logistic Regression, RF, and Lasso can be attributed to several factors. Logistic Regression may struggle to capture the

Table 5 ACC of AGECE compared with different regression models

	Logistic regression	Random forest	Lasso	AGECE
ACC	0.5924	0.5421	0.5601	0.7250

The value in bold font symbolizes the best performance

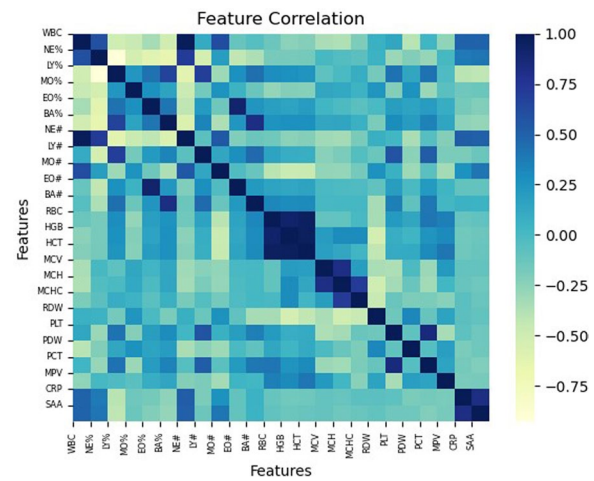


Fig. 3 A heatmap visualization of the correlation between features

complex non-linear relationships present in the data, leading to suboptimal predictive performance. RF, while robust in certain contexts, may face challenges in handling the specific characteristics of the asthma prediction task. Lasso, being a feature selection method, may not effectively discern the important features contributing to asthma prediction, resulting in reduced accuracy.

Impact of different subsets of features on the effectiveness of AGECE

This experiment aimed to determine the discriminability of various feature sets in asthma prediction. Here, we explored three groups of features. The first set of features was extracted by considering prior knowledge from relevant medical literature, such as [37, 38], yielding a group consisting of 14 key features. The characteristics of these features are described as follows: WBC,

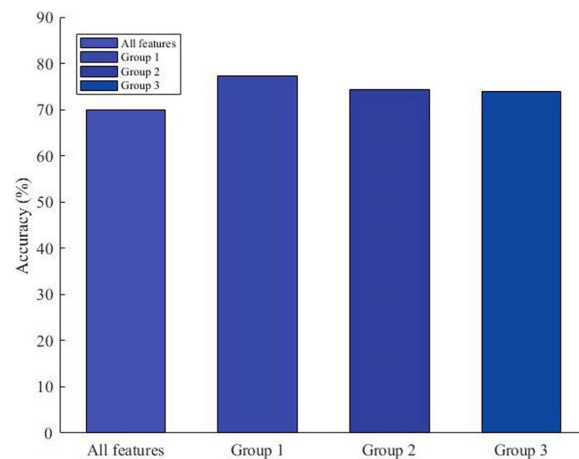


Fig. 4 The ACC of AGECE on different set of features

LY%, MO%, LY#, MO#, EO#, BA#, RBC, MCH, MCHC, RDW, PLT, PDW, MPV. Based on this, we further investigated the correlation between features using a heat map. As may be noticed in Fig. 3, we observed that PDW and MPV among the indicators of blood routine have a great impact on the final results, so we take these two indicators as the center. Then, the heat map was used to find the features that are highly correlated with those two indicators, leading to two additional sets of features. Thus, the second group has 13 features: PDW, MPV, RDW, BA%, EO%, MO%, LY#, PCT, PLT, MCV, HCT, HGB, and RBC. The third group has 15 features, which are shown as follows. MPV, PCT, PDW, RDW, MCH, MCV, HCT, HGB, RBC, BA#, EO#, BA%, EO%, LY%, and NE%.

According to the results in Fig. 4, our model obtained an accuracy of 78.18%, 75.29% and 72.92% under the first, second and third groups, respectively. Notably, AGECD demonstrated the highest accuracy (78.18%) in the first set, indicating that the selected features were particularly effective in distinguishing between groups. In contrast, the model achieved slightly lower accuracies of 75.29% and 72.92%, respectively, for the second and third sets, suggesting that some of the features employed in these sets were not as discriminatory. Interestingly, it can be observed that the third group, despite having more features (15), did not outperform the second group (13 features), meaning that the additional features may not have significantly contributed to the classification task. These findings thus underscore the fact that not all added features would necessarily improve the performance of a classification model.

As a result of the above, we further conducted experiments on each of the 24 features to determine which input features are most salient. Based on these

experiments, we present a graphical representation of the performance of the classification model using a ROC curve. This plots the true positive rate (TPR), also known as sensitivity, against the false positive rate (FPR), also referred to as specificity. As shown in Fig. 5, we only display results of MPV, LY% and RDW with more obvious effects. Accordingly, it can be observed that the curve area formed by these three indicators is greater than $y = x$, meaning that our model has practical significance in the three indicators. At the same time, it can also be observed that MPV has a better effect on the classification of asthma compared to other indicators.

Discussion

In this study, we presented a novel model for asthma prediction that incorporates an affinity graph enhanced classifier and utilized previously unexplored clinical indicators. This combination sets our study apart from previous works, offering distinct advantages and contributing to the field of medical predictive modeling.

One of the key advantages of our approach was the integration of affinity graph to capture correlations between samples. This aspect of our approach enhanced the ability of our model to capture intricate interactions within the data and improve overall prediction performance. In addition to the use of the affinity graph, our study focused on utilizing unique clinical predictors for asthma prediction. We extracted 24 clinical indicators, including blood count differentials and red blood cell indices. As far as we know, the selected predictors have not been previously utilized “solely” for the training of ML models in the context of asthma prediction. This inclusion thus expands the scope of predictors used in asthma prediction models and can potentially uncover new insights into the disease. Moreover, our study demonstrated that utilizing these unique clinical predictors alone can achieve competitive performance, with an ACC of 72.50% and an AUC of 74.01%, as shown in Tables 3 and 4, respectively. This highlights the effectiveness of our proposed model, showing that the employed clinical indicators can provide meaningful and discriminative information for asthma prediction. Furthermore, the use of these clinical predictors offers advantages in terms of simplicity, interpretability and generalization. For example, collecting and integrating various data sources can be challenging and time-consuming, whereas our approach simplifies the prediction process by focusing exclusively on clinical data, which are often readily available in medical settings. This streamlined approach enhanced the ease of implementation, and, we hope that the clinical community may cautiously consider the adoption of our model to facilitate prompt detection and management of asthma to avoid exacerbations and hospitalizations.

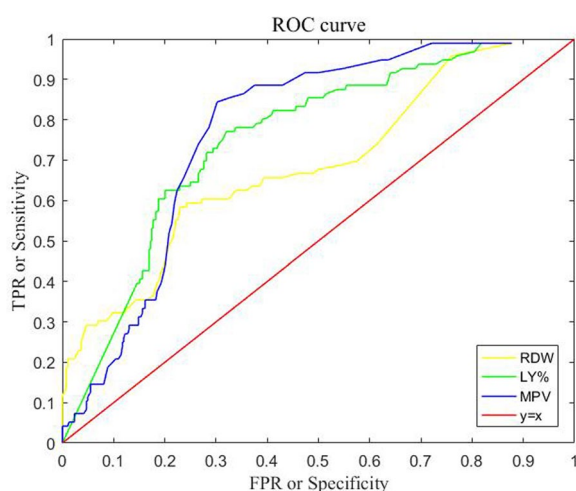


Fig. 5 The ROC curve of the true positive rate against the false positive rate with respect to MPV, LY% and RDW indicators

In addition to the improvements offered from the above two perspective, it is noteworthy to highlight the robustness of our approach. While previous studies, such as [9, 39], have often relied on traditional ML algorithms and utilized data from multiple sources, such as age, gender, lung function measurements, and medical history to make predictions, our study demonstrated that using a focused set of unique clinical predictors can achieve a comparable or even superior performance if the predictive model can capture the correlation between samples. To further emphasize this, we examined the results reported in existing literature for asthma prediction models. While the specific studies may vary, a comprehensive review of recent works [40] revealed that the performance accuracy of most asthma predictive models is generally >65%. In comparison, our study achieved an accuracy of 70% using only the selected clinical predictors.

Furthermore, based on the evaluation of the effect of three different subsets of features on the performance of AGECE, we found that the accuracy of the proposed model can reach 78.18%, with the accuracy across all three sets ranging from 72.92% to 78.18%. This variation underpins the importance of feature selection in enhancing the performance of classification models. More specifically and consistent with previous medical studies by Panet al. [37] and Zhu et al. [38], the first group, with its specific set of features, demonstrated the highest accuracy. This suggests that the co-existence of certain indicators, such as WBC, LY%, MO%, LY#, MO#, EO#, BA#, RBC, MCH, MCHC, RDW, PLT, PDW, and MPV, can play a crucial role in distinguishing asthma cases. Besides, the observed significance of MPV in our study suggests that platelet-related factors may play a role in diagnosing asthma. This finding aligns with emerging evidence in [41–43] that implicates platelet activation and inflammation in the pathogenesis of asthma. Additionally, the differential impact of LY% and RDW on asthma classification underscores the intricate interplay between lymphocyte percentages and red cell distribution width in the context of asthma-related processes. These insights provide a foundation for exploring potential biomarkers related to immune response and erythropoiesis in asthma. Therefore, it is also hoped that this knowledge will further guide clinicians in prioritizing these indicators for prompt and accurate diagnosis of asthma, ultimately reducing the burden on healthcare systems. Another advantage of our approach is its potential for easy extension to other diseases detection. This flexibility demonstrates the broader applicability and impact of our study. Nonetheless, even though the proposed approach has been validated to be effective, our study may have been limited by the size of the dataset. Although we tried

to mitigate such effects via the incorporation of dimensionality reduction in our model, we believe that, in the future, the accuracy of AGECE can be further improved by increasing the sample population. Moreover, recent studies such as [44], have found that the level of heavy metals in serum was higher in individuals with acute exacerbation of Chronic Obstructive Lung Disease (COPD); therefore, in future work, we hope to employ a combination of these features with the other blood markers used in this study to enhance accuracy.

Conclusions

In this paper, we proposed a new method for predicting asthma using an affinity graph enhanced classifier. Our approach specifically addressed the limitation of existing models in terms of capturing the correlation between data samples. As a result, the accuracy of our model was improved in asthma prediction. This was accomplished by utilizing domain knowledge through the affinity graph. Compared with existing state-of-the-art related models concerning ACC and AUC, our AGECE demonstrated significant improvement in asthma prediction. To the best of our knowledge, this is the first study that directly exploits the affinity graph for classification tasks, and the results have shown its effectiveness. In addition, the proposed approach is completely data-driven and can easily be generalized to other prediction tasks, thus providing a framework for future research. Moreover, beyond the immediate scope of asthma prediction, the implications of our findings extend to the broader context of asthma management and healthcare. The enhanced accuracy and novel methodology introduced by AGECE holds potential benefits for improving early asthma detection, thus enabling more proactive and targeted interventions. This, in turn, could contribute to the optimization of patient care, reduction of healthcare costs, and the overall enhancement of asthma management strategies.

Author contributions

Conceptualization and supervision: YC and LS; methodology: DL and SEA; validation: DZ and XS; review and editing: XS and YC; funding acquisition: YC. All the authors have read and agreed to the published version of the manuscript.

Funding

His study was supported by the Top Talents Project of the Wuxi Taihu Lake Talent Plan (2020THRC-GD-7), the 333 project of Jiangsu Province in 2022 (ZUZHIBU 202221001), and "Light of Taihu Lake" scientific and technological breakthrough from Wuxi science and Technology Bureau (R & D of medical and health technology, Y20212006).

Availability of data and materials

The authors declare that all data supporting the findings of this study are available withing the article and its additional files or by contacting the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

The experimental protocol was established according to the ethical guidelines of the Declaration of Helsinki and was approved by the Ethic Committee of the Affiliated Shuguang Hospital of Shanghai Traditional Chinese Medicine University.

Consent for publication

All the authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare that they have no competing financial interests.

Author details

¹Present Address: Department of Respiratory, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi 214023, China. ²School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China. ³Department of Computer Science, University of Nigeria, Nsukka, Nigeria. ⁴Clinical Research Center, The Affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi 214023, China. ⁵Department of Clinical Laboratory, Shuguang Hospital Affiliated to Shanghai University of Chinese Traditional Medicine, Shanghai 201203, China.

Received: 14 October 2023 Accepted: 6 January 2024

Published online: 24 January 2024

References

- Khurana S, Jarjour NN. Systematic approach to asthma of varying severity. *Clin Chest Med*. 2019;40(1):59–70.
- Talha SAAM, MagzoubAlhaj OS, Elhag A. Evaluation of asthma control assessment in school-age asthmatic children. *J Med-Clin Res Rev*. 2022;6:1–6.
- Zampogna E, Zappa M, Spanevello A, Visca D. Pulmonary rehabilitation and asthma. *Front Pharmacol*. 2020;11:542.
- Padem N, Saltoun C. Classification of asthma. *Allergy Asthma Proc*. 2019;40(6):385–8.
- Busse WW, Wenzel SE, Casale TB, FitzGerald JM, Rice MS, Daizadeh N, Deniz Y, Patel N, Harel S, Rowe PJ, et al. Baseline FeNO as a prognostic biomarker for subsequent severe asthma exacerbations in patients with uncontrolled, moderate-to-severe asthma receiving placebo in the LIBERTY ASTHMA QUEST study: a post-hoc analysis. *Lancet Respir Med*. 2021;9(10):1165–73.
- Moral L, Vizmanos G, Torres-Borrego J, Praena-Crespo M, Tortajada-Girbés M, Pellegrini FJ, Asensio Ó. Asthma diagnosis in infants and preschool children: a systematic review of clinical guidelines. *Allergol Immunopathol (Madr)*. 2019;47(2):107–21.
- Xiang Y, Ji H, Zhou Y, Li F, Du J, Rasmay L, Wu S, Zheng WJ, Xu H, Zhi D, et al. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: retrospective cohort study. *J Med Internet Res*. 2020;22(7):e16981.
- Ross MK, Yoon J, van der Schaar A, van der Schaar M. Discovering pediatric asthma phenotypes on the basis of response to controller medication using machine learning. *Ann Am Thorac Soc*. 2018;15(1):49–58.
- Kothalawala DM, Murray CS, Simpson A, Custovic A, Tapper WJ, Arshad SH, Holloway JW, Rezwan FI, et al. Development of childhood asthma prediction models using machine learning approaches. *Clin Transl Allergy*. 2021;11(9):e12076.
- Spathis D, Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics J*. 2019;25(3):811–27.
- Rani A, Sehwat H. Role of machine learning and random forest in accuracy enhancement during asthma prediction. In: 2022 10th international conference on reliability, infocom technologies and optimization (trends and future Directions) (ICRITO) IEEE. pp. 1–10.
- Ekpo RH, Osamor VC, Azeta AA, Ikekana E, Amos BO. Machine learning classification approach for asthma prediction models in children. *Health Technol*. 2023;13:1–10.
- Zein JG, Wu CP, Attaway AH, Zhang P, Nazha A. Novel machine learning can predict acute asthma exacerbation. *Chest*. 2021;159(5):1747–57.
- Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81–106.
- Statistics L B, Breiman L. Random Forests. *Mach Learn*. 2001:5–32.
- Hearst MA, Dumais ST, Osman E, et al. Support vector machines. *IEEE Intell Syst App*. 1998;13(4):18–28.
- Borrie K. Neural networks and their applications. *Rev Sci Instrum*. 1994;65(6):1803–32.
- Heckerman, Wellman MP. Bayesian networks. *CACM*. 1995;38:27–31.
- He X. Locality preserving projections. *NIPS*. 2003;16(1):186–97.
- He X, Cai D, Yan S, Zhang, HJ. Neighborhood preserving embedding. In: Tenth IEEE international conference on computer vision (ICCV'05) Volume 1 IEEE 2, pp. 1208–13.
- Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y. Robust recovery of subspace structures by low-rank representation. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(1):171–84.
- Elhamifar E, Vidal R. Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(1):2765–81.
- Xie X, Guo X, Liu G, Wang J. Implicit block diagonal low-rank representation. *IEEE Trans Image Process*. 2018;27(1):477–89.
- Lu C, Feng J, Lin Z, Mei T, Yan S. Subspace clustering by block diagonal representation. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(2):487–501.
- Shao R, Xu N, Geng X. Multi-label learning with label enhancement. In: 2018 IEEE international conference on data mining (ICDM). IEEE. pp. 437–46.
- Wang ZY, Abhadiomhen SE, Liu ZF, Shen XJ, Gao WY, Li SY. Multi-view intrinsic low-rank representation for robust face recognition and clustering. *IET Image Proc*. 2021;14:15.
- Abhadiomhen SE, Wang Z, Shen X. Coupled low rank representation and subspace clustering. *Appl Intell*. 2022;52:530–46.
- Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv 2010; 1009: 5055*.
- Gordon G, Tibshirani R. Karush-kuhn-tucker conditions. *Optimization*. 2012;10:725.
- Nie F, Wang X, Jordan M, Huang H. The constrained laplacian rank algorithm for graph-based clustering. *Proc AAAI Conf Artif Intell*. 2016;30:1969–76.
- Zhang QW, Zhong Y, Zhang ML. Feature-induced labeling information enrichment for multi-label learning. In: Proceedings of the AAAI conference on artificial intelligence. 2018. p. 32.
- Guo X, Wang XB, Ling H. Exclusivity regularized machine: a new ensemble SVM classifier. *IJCAI*. 2017. p. 36.
- Zhang F, Odonnell LJ. Support vector regression. In: Machine learning. Academic Press; 2020. p. 123–40.
- Gu X, Angelov PP. Multiclass Fuzzily Weighted Adaptive-Boosting-Based Self-Organizing Fuzzy Inference Ensemble Systems for Classification. *IEEE T Fuzzy Syst*. 2021; 30: 3722–35.
- Gao T, Yang J, Jiang S. A novel incipient fault diagnosis method for analog circuits based on GMKL-SVM and wavelet fusion features. *IEEE Trans Pattern Anal Mach Intell*. 2020;70:1–15.
- Xing HJ, Liu WT. Robust AdaBoost based ensemble of one-class support vector machines. *Inform Fusion*. 2020;55:45–58.
- Pan R, Ren Y, Li Q, Zhu X, Zhang J, Cui Y, Yin H. Neutrophil-lymphocyte ratios in blood to distinguish children with asthma exacerbation from healthy subjects. *Int J Immunopathol Pharmacol*. 2023;37:3946320221149849.
- Zhu X, Song H, Chen Y, Han F, Wang Q, Cui Y. Neutrophil-to-Lymphocyte ratio and platelet-to-lymphocyte ratio in blood to distinguish lung cancer patients from healthy subjects. *Dis Mark*. 2020;2020:8844698.
- Finkelstein J, Jeong IC. Machine learning approaches to personalize early prediction of asthma exacerbations. *Ann N Y Acad Sci*. 2017;1387(1):153–65.
- Kothalawala DM, Kadalayil L, Weiss VBN, Kyyaly MA, Arshad SH, Holloway JW, Rezwan FI. Prediction models for childhood asthma: a systematic review. *Pediatr Allergy Immunol*. 2020;31(6):616–27.

41. Takeda T, Morita H, Saito H, Matsumoto K, Matsuda A. Recent advances in understanding the roles of blood platelets in the pathogenesis of allergic inflammation and bronchial asthma. *Allergol Int.* 2018;67(3):326–33.
42. Pitchford S, Cleary S, Arkless K, Amison R. Pharmacological strategies for targeting platelet activation in asthma. *Curr Opin Pharmacol.* 2019;46:55–64.
43. Luo L, Zhang J, Lee J, Tao A. Platelets, not an insignificant player in development of allergic asthma. *Cells.* 2021;10(8):2038.
44. Albayrak L, Türksoy VA, Khalilov R, Eftekhari A. Investigation of heavy metal exposure and trace element levels in acute exacerbation of COPD. *J King Saud Univ Sci.* 2023;35(1): 102422.

Publisher's Note

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