BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Risk prediction of hyperuricemia based on particle swarm fusion machine learning solely dependent on routine blood tests

Min Fang[1,2], Chengjie Pan[1†], Xiaoyi Yu[3], Wenjuan Li[1,2,5*], Ben Wang[1,2], Huajian Zhou[4], Zhenying Xu[4] and Genyuan Yang[1]

Abstract

Hyperuricemia has seen a continuous increase in incidence and a trend towards younger patients in recent years, posing a serious threat to human health and highlighting the urgency of using technological means for disease risk prediction. Existing risk prediction models for hyperuricemia typically include two major categories of indicators: routine blood tests and biochemical tests. The potential of using routine blood tests alone for prediction has not yet been explored. Therefore, this paper proposes a hyperuricemia risk prediction model that integrates Particle Swarm Optimization (PSO) with machine learning, which can accurately assess the risk of hyperuricemia by relying solely on routine blood data. In addition, an interpretability method based on Explainable Artificial Intelligence(XAI) is introduced to help medical staff and patients understand how the model makes decisions. This paper uses Cohen's d value to compare the differences in indicators between hyperuricemia and non-hyperuricemia patients and identifies risk factors through multivariate logistic regression. Subsequently, a risk prediction model is constructed based on the parameter optimization of five machine learning models using the PSO algorithm. The accuracy and sensitivity of the proposed particle swarm fusion Stacking model reach 97.8% and 97.6%, marking an improvement in accuracy of over 11% compared to the state-of-the-art models. Finally, a sensitivity analysis of factors affecting the prediction results is conducted using the XAI method. This paper has also developed a Health Portrait Platform that integrates the proposed risk prediction models, enabling real-time online health risk assessment. Since only routine blood test data are used, the new model has better feasibility and scalability, providing a valuable reference for assessing the risk of hyperuricemia occurrence.

Keywords Disease risk prediction, Hyperuricemia, Particle swarm fusion machine learning, Routine blood test, Model explainability

†Chengjie Pan contributed equally to this work.

Min Fang and Chengjie Pan: co-first author.

*Correspondence:
Wenjuan Li
liwenjuan@hznu.edu.cn
[1]School of Information Science and Technology, Hangzhou Normal University, Yuhangtang Rd., Hangzhou, Zhejiang 311121, China
[2]Engineering Research Center of Mobile Health Management System, Ministry of Education, Yuhangtang Rd., Hangzhou, Zhejiang 311121, China
[3]College of Engineering, Zhejiang University, Yuhangtang Rd., Hangzhou, Zhejiang 310058, China
[4]Zhejiang Yishan Smart Medical Research Co., Ltd., Liangmu Rd., Hangzhou, Zhejiang 311121, China
[5]Computer Science and Technology, Zhejiang University, Yuhangtang Rd., Hangzhou, Zhejiang 310058, China

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Uric acid is the end product of purine catabolism in the human body and plays a positive role in antioxidant activity and blood pressure maintenance [1]. Uric acid is closely related to human health conditions, and imbalances in uric acid can lead to a series of adverse reactions [2, 3]. Under normal circumstances, about 70% of the uric acid produced in the body is excreted through the kidneys in urine, while the remaining 30% is excreted through the intestines [4]. Overproduction of uric acid or insufficient excretion leads to higher than normal uric acid levels in the blood, a condition known as hyperuricemia [5].

The global prevalence of hyperuricemia varies significantly across different regions, but with the development of the social economy and lifestyle changes, the prevalence is generally on the rise. In Asia, particularly in China and Japan, the prevalence of hyperuricemia and its complications is high and increasingly tends to be younger. According to the data from the 2018–2019 China Chronic Disease and Risk Factor Surveillance, the prevalence of hyperuricemia among Chinese adult residents is 14.0%, with male and female prevalence rates being 24.5% and 3.6%, respectively [6–8].

Hyperuricemia has a very high correlation with gout, and clinical evidence suggests that the former is an important biochemical basis for gout [9]. In addition, the persistent increase in blood uric acid levels is also associated with the occurrence and development of kidney diseases, endocrine metabolic diseases, cardiovascular and cerebrovascular diseases, etc. [10]. Hyperuricemia is also an independent risk factor for chronic kidney disease, hypertension, cardiovascular and cerebrovascular diseases, and diabetes, and is an independent predictor of premature death [11]. It is evident that hyperuricemia poses a significant threat to human health, increasing the risk of various diseases, and has become a serious public health issue.

However, hyperuricemia is not easily detected in its early stages, mainly because it has no obvious symptoms [12]. Therefore, most patients rarely notice any abnormalities in their blood uric acid levels before gout or other serious complications. Moreover, the gradual increase in uric acid levels is usually slow, and slight early increases can easily be overlooked. Consequently, without regular health check-ups, especially blood tests targeting uric acid levels, hyperuricemia may go undiagnosed for a long period.

Considering the high cost and long duration of diagnosing hyperuricemia in large populations, developing a predictive model to screen high-risk groups and reduce the screening scope is an effective alternative approach. Currently, researchers mainly use machine learning methods to construct risk prediction models for hyperuricemia. For instance, literature [13] built a hyperuricemia risk prediction model through stepwise logistic regression analysis, decision tree algorithms, and Lasso regression analysis, and these models have shown good predictive performance on both the training and validation sets. Shi Jiacheng and others developed a hyperuricemia prediction model using four data items, but its ROC curve was only 0.745 [14]. Literature [15] provided new quantitative markers for the early detection and prognosis prediction of hyperuricemia using a stacked multimodal machine learning model. This model has shown satisfactory performance on the training set, internal test set, and external test set. Wang YJ and others applied artificial neural network algorithms to the construction of hyperuricemia models, achieving relatively ideal predictive results [16]. Literature [9] used the Extreme Gradient Boosting (XGBoost) algorithm to establish a model predicting the risk of hyperuricemia in people taking low-dose aspirin, demonstrating good predictive accuracy.

The aforementioned studies have demonstrated the immense potential of machine learning in disease risk prediction, but they also have limitations. For instance, due to data quality issues, datasets may be biased, leading to lower model accuracy. Additionally, predictive models trained with deep learning algorithms require significant resource demands and energy consumption. Moreover, most current risk prediction models for hyperuricemia include both complete blood routine and biochemical indicators as input features. Biochemical tests are more time-consuming and expensive, whereas blood routine tests are a more common and widely applied method of testing that can be easily conducted in all medical institutions [17]. This suggests that the potential for prediction using only blood routine data has not yet been fully explored. Therefore, this study constructed a hyperuricemia risk prediction model that utilizes machine learning algorithms and relies solely on routine blood test data, incorporating PSO for autonomous optimization of the machine learning models [18]. The constructed model can quickly and accurately assess the risk of hyperuricemia based solely on routine blood test data. Additionally, the XAI, an interpretable machine learning technology, was introduced to conduct a sensitivity analysis of the influencing factors in the multivariate prediction of hyperuricemia, revealing the key influencing factors of hyperuricemia [19].

The main contributions of this paper are as follows.

- For the risk prediction of hyperuricemia, this paper has conducted numerous comparative experiments using benchmark and ensemble models and ultimately proposes a fusion model based on PSO and ensemble learning [20]. The proposed model shows a good performance on real datasets, with an

11% increase in predictive accuracy compared to the latest models.
- The proposed model relies solely on routine blood test data, reducing detection costs and shortening the detection cycle, providing a new technological means for the timely diagnosis of high-risk populations for hyperuricemia.
- Real datasets (including physical examination and diagnostic data) were used to implement modeling and testing of the prediction model. At the same time, explainable artificial intelligence methods, including SHAP and LIME, were introduced to enhance the transparency and credibility of the prediction model, making it more persuasive.
- This paper has developed a health portrait platform that integrates the proposed disease risk prediction model, achieving real-time online health risk assessment.

Related works

Disease prediction helps to achieve early intervention of diseases and control their progression, possessing high application value. As a result, it has garnered widespread attention from researchers and has yielded a plethora of research outcomes [21, 22].

Traditional disease risk prediction is mainly based on the Cox proportional hazards regression model. For example, Wang et al. establish a risk prediction model based on the Cox model for stroke and death in atrial fibrillation patients based on the Framingham Heart Study [23]. The H-L statistics of the stroke prediction model and the stroke or death prediction model are 7.6 and 6.5, respectively, with the AUC of the stroke prediction model being 0.66 and the AUC of the stroke or death prediction model being 0.70. Khosla et al. used feature selection and machine learning methods to predict the incidence of stroke within 5 years [24]. Using L1 regularized logistic regression for feature selection and Support Vector Machine (SVM) for prediction, the average test AUC using 10-fold cross-validation is 0.764, which is better than the L1 regularized Cox model. Kun Lv and colleagues constructed diabetes risk prediction models using various machine learning methods and ultimately found that the model based on logistic regression performed the best [25]. On the validation set, the model achieved an AUC of 0.899, a sensitivity of 0.850, and a specificity of 0.811. Through the analysis of feature importance, it was revealed that BMI, age, and gender are important factors in the risk stratification of diabetes.

In the field of hyperuricemia prediction, researchers have also proposed several valuable reference models. For example, Lee et al. explore machine learning methods for hyperuricemia prediction models based on basic health check test results [26]. Under the maximum sensitivity criterion, the naive Bayes (NB) algorithm exhibits the highest sensitivity (0.73), followed by the Random Forest (RF) algorithm (0.66); Under the maximum balanced classification rate (BCR) standard, the RF algorithm exhibits the highest BCR (0.68), followed by the NB algorithm (0.66). Compared with the traditional logistic regression model (AUC = 0.568), the NB (AUC = 0.669) and RF (AUC = 0.775) models showed significantly better predictive performance ($p < 0.001$). Hou et al. constructed a hyperuricemia prediction model to assist in early prevention and screening [4]. The study analyzed the risk factors of gender and age groups through Pearson correlation analysis, binary logistic regression, and ROC curve analysis. The results showed that total protein (TP), low-density lipoprotein cholesterol (LDL-C), and glucose (GLU) were risk factors for hyperuricemia, and the AUC value of the SVM-based model on the validation set was 0.875. Zheng et al. developed a predictive model using an occupational health examination dataset to predict the risk of hyperuricemia in steelworkers [27]. The study used three models: logistic regression, convolutional neural network (CNN), and XG Boost, and selected six influencing factors through LASSO regression. Ultimately, the XG Boost model performed the best in discrimination, calibration, and clinical applicability. Based on dietary information, the literature [28] uses logistic regression models to screen for dietary risk factors related to hyperuricemia. Then, the Artificial Neural Network (ANN) is used to construct a prediction model, and the accuracy of the model is evaluated through ROC curve analysis. The results showed that the area under the ROC curve of the ANN model on the training set and validation set was 0.827 and 0.814, respectively.

Since traditional machine learning methods do not perform satisfactorily in terms of accuracy for disease prediction, many scholars have further proposed integrated methods to implement further optimization. Mahajan et al. studied the application of ensemble learning in disease prediction [29]. The study reviewed 45 articles published between 2016 and 2023 and found that although ensemble learning methods (including Bagging, Boosting, Stacking, and Voting) are widely used in disease prediction, the Stacking method performs the best in accuracy, however, with the lower frequency of usage than Bagging and Boosting. Zhang et al. developed an ensemble model for hyperuricemia based on a prospective health examination population [30]. The research results indicate that 15 important features were selected from 23 clinical variables. The AUC of the stacked ensemble model is 0.854, which is superior to the other three models (Support Vector Machine, Decision Tree C5.0, and XGBoost, with AUC of 0.848, 0.851, and 0.849, respectively).

The aforementioned studies have certain limitations when using machine learning or integrated models for

disease prediction. These mainly include: 1) Not using real datasets for modeling, which leads to models that cannot be used in practical applications or have insufficient prediction accuracy; 2) Low data quality or obvious biases in the datasets resulting in low model accuracy; 3) The vast majority of hyperuricemia prediction models require a large number of detection indicators, including some biochemical indicators, leading to high costs and long prediction cycles. In contrast, this paper uses real detection data from hospitals and achieves data balance through technical means. It relies solely on routine blood test indicators, making the method simple and easy to promote, and it improves prediction accuracy through integrated learning.

Experimental framework

Overview of the technical framework

Figure 1 shows the overview of the technical framework. First, the collected data is preprocessed [31]. To address the issue of class imbalance in the dataset, we used oversampling methods on the dataset. The dataset is divided into a training set and a test set in a ratio of 7:3 (random_state = 42) [32]. The training set is used for 10-fold cross validation (random_state = 0) training of the model, and the test set is used for the final model evaluation [33, 34]. To filter features, a univariate analysis of factors affecting hyperuricemia is performed, and factors with statistical significance ($P < 0.05$) from the univariate analysis are used as independent variables, with the occurrence of hyperuricemia as the dependent variable for Logistic regression multivariate analysis [35]. For NULL values in the independent variables, we used multiple imputations to fill in the missing values. Next, we conducted a comparison on several baseline machine learning methods, including Logistic Regression (LR), RF, SVM, Deep Neural Network (DNN), and XGBoost and utilized the PSO algorithm to optimize the parameters [36, 37]. Through performance evaluation, the best baseline model was

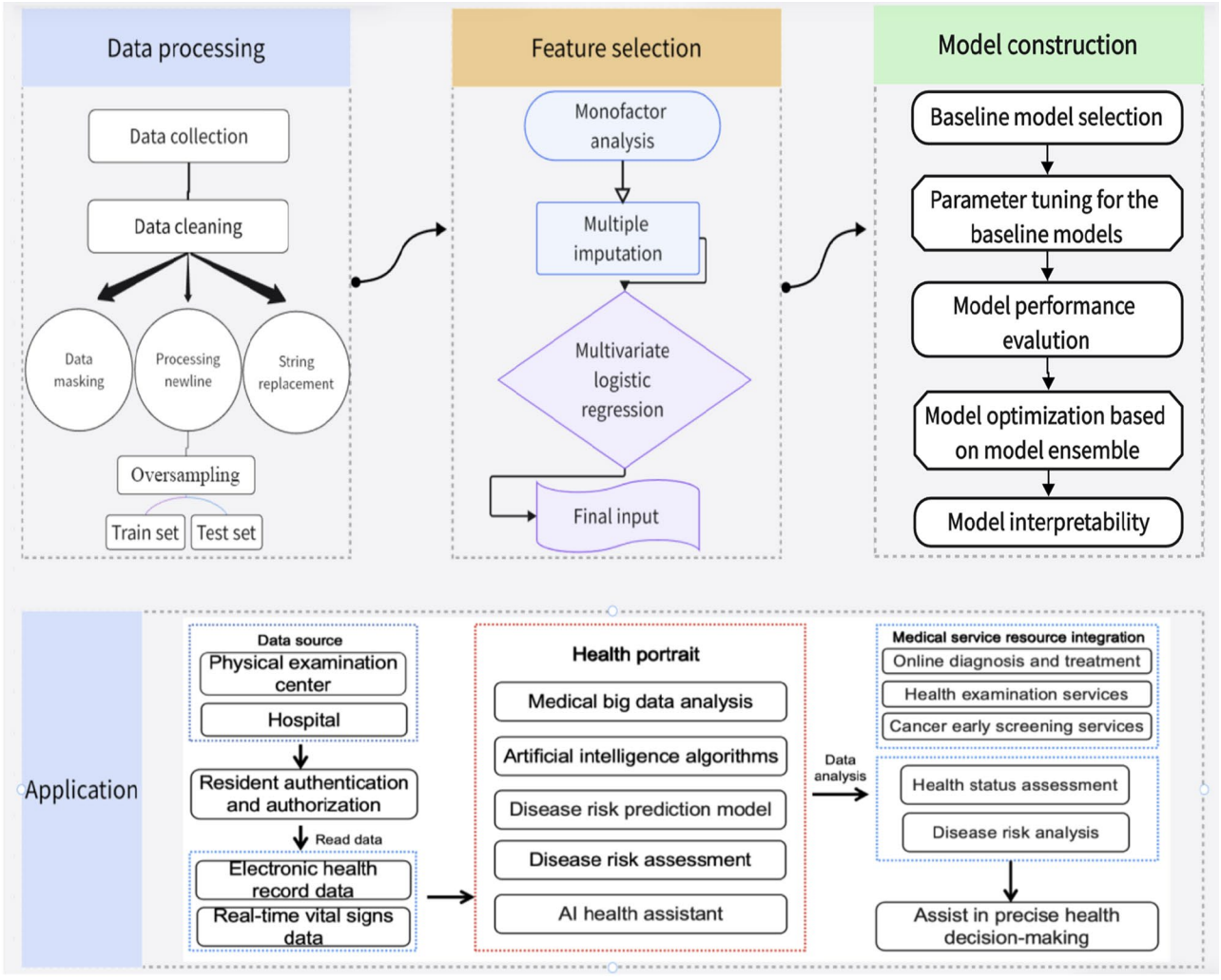

Fig. 1 Overview of the technical framework

selected, and the accuracy of the proposed model was enhanced by the model ensemble method. Finally, by introducing the XAI, we explained the relationship between each feature and the risk of disease, which helps to better understand how the model makes decisions [38, 39]. Based on model validation and optimization, we developed a health portrait platform, which is equipped with the predictive algorithms proposed in this paper [40]. The platform can manage the health data of registered users in a unified manner and make disease risk predictions.

DataSet processing

The data used in this study is sourced from hospitals in two regions of Zhejiang. The data is categorized and stored in two separate tables, with the physical examination table containing a total of 30,606,676 physical examination records and the disease table containing a total of 1,958,752 disease records. The dataset includes 26 routine blood test features, specifically: White Blood Cells (WBC), Red Blood Cells (RBC), Hemoglobin (HGB), Packed Cell Volume (PCV), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Platelet Count (PLT), Lymphocyte Percentage (LYM), Neutrophil Percentage (NEUT%), Monocyte Percentage (MONO%), Basophil Percentage (BASO%), Eosinophil Percentage (EOS%), Lymphocyte Absolute (LYM#), Neutrophil Absolute (ANC), Monocyte Absolute (MONO#), Basophil Absolute (BASO#), Eosinophil Absolute (EOS#), Red Cell Distribution Width - CV (RDW-CV), Red Cell Distribution Width Standard Deviation (RDW-SD), Platelet Distribution Width (PDW), Mean Platelet Volume (MPV), Platelet Crit (PCT), Age (AGE), Gender (SEX), and Weight (WEIGHT). Table 1 shows the relevant features of the data in this paper.

To better construct the predictive model, it is necessary to preprocess the dataset. In this study, the first step is to complete the data cleaning by removing sensitive data and newline characters and replacing parentheses. Then it uses the pandas.merge and pivot_table functions in Python to join the two tables and obtain a pivot table.

After preprocessing, a total of 144 cases with hyperuricemia and 6,271 cases without hyperuricemia were obtained.

Table 1 Overview of data related features

Blood routine indicators	WBC, RBC, PCV, NEUT%, ANC, MONO%, MONO#, BASO%, BASO#, EOS%, EOS#, MCV, MCHC, MCH, LYM, LYM#, RDW-CV, RDW-SD, PDW, PCT, MPV, PLT, HGB
Sex	Male (2824), Female (3591)
Age	0–18 (9), 19–35 (587), 36–60 (2190), 60+ (3629)
Hyperuricemia	Yes (144), No (6271)

Due to the severe class imbalance in the dataset, this paper proposed a SMOTE-based oversampling method to make the number of positive and negative samples equal. The SMOTE-based data processing algorithm is shown in Algorithm 1.

Algorithm 1 SMOTE-based Data Processing Algorithm

1: **Input:** Medical Examination Data Sheet T_1, Patient Information Form T_2
2: **Output:** Balanced data S that can be input into the model
3: Read in the data tables T_1 and T_2 and perform some preprocessing
4: Merge T_1 and T_2 on common keys
5: Specify row index name and column index name to obtain pivot table
6: Filter pivot table to select relevant blood routine columns
7: Apply multiple interpolation methods to fill in missing values
8: Apply SMOTE to handle imbalanced datasets
9: Separate the processed data into features X and labels Y
10: **Return** balanced data S (X, Y)

Feature analysis

This paper selects the t-test method to perform univariate analysis on the dataset to identify factors affecting hyperuricemia [41, 42]. The results of the t-test can determine whether there are significant differences between the data (through P-values), but it cannot explain the actual magnitude of the differences. Furthermore, the test results are easily influenced by the sample size: a smaller sample may lead to insignificant results, while a larger sample may amplify minor differences. Cohen's d is a standard- ized effect size that is not directly affected by sample size and can better reflect the actual size of the differences. Therefore, this paper combines Cohen's d values to rep- resent the differences in physical examination indicators between the hyperuricemia group and the non-hyperuricemia group. And its definition is shown in Eq. (1):

$$Cohen's = \frac{(x_1 - x_2)}{s} \tag{1}$$

Where x_1 and x_2 are the means of the two populations, and s is the standard deviation of the population. From the analysis results of Table 2, it can be seen that the characteristics of NEUT%, ANC, WEIGHT, MONO#, BASO#, EOS#, LYM, RDW-CV, HGB, AGE, and SEX are statistically significant ($P < 0.05$) in the comparison between the two categories. The top 5 features with significant differences are WEIGHT, SEX, BA#, ANC, and

Table 2 Univariate analysis results of factors affecting hyperuricemia

Indicators	Hyperuricemia	Non-hyperuricemia	*P*	Cohen_d	t
WBC	5.93 ± 1.25	5.83 ± 1.66	0.744	0.061	0.327
RBC	4.62 ± 0.69	4.55 ± 0.52	0.121	0.131	1.548
PCV	41.16 ± 5.21	42.41 ± 4.12	0.108	0.302	−1.608
NEUT%	63.29(53.93,72.65)	61.01(52.06,69.96)	< 0.05	0.254	2.974
ANC	4.36(2.63,6.09)	3.71(2.34,5.08)	< 0.05	0.47	5.568
WEIGHT	70.26(58.33,82.19)	60.79(49.27,72.31)	< 0.05	0.821	7.693
MONO%	5.75(3.68,7.82)	5.75(3.79,7.71)	0.962	0	−0.047
MONO#	0.38(0.22,0.54)	0.34(0.20,0.48)	< 0.05	0.284	3.354
BASO%	0.42 ± 0.19	0.38 ± 0.28	0.193	0.144	1.302
BASO#	0.03 ± 0.01	0.02 ± 0.02	< 0.05	0.506	2.302
EOS%	2.43 ± 1.81	2.32 ± 2.04	0.535	0.054	0.621
EOS#	0.16 ± 0.13	0.14 ± 0.13	< 0.05	0.154	1.991
MCV	93.19 ± 5.97	92.90 ± 5.65	0.558	0.051	0.585
MCHC	329.01(317.79,340.23)	332.25(323.83,340.67)	0.229	0.38	1.203
MCH	30.65 ± 2.27	30.46 ± 2.15	0.298	0.088	1.04
LYM	28.16(19.75,36.57)	30.57(22.14,39.00)	< 0.05	0.286	−3.35
LYM#	1.84(1.24,2.44)	1.80(1.14,2.46)	0.476	0.061	0.712
RDW-CV	28.24 ± 16.26	22.54 ± 14.89	< 0.05	0.381	4.131
RDW-SD	43.02 ± 8.71	41.23 ± 10.44	0.051	0.172	1.951
PDW	16.04 ± 0.92	15.98 ± 1.32	0.604	0.046	0.519
PCT	0.20 ± 0.06	0.20 ± 0.06	0.997	0	0.004
MPV	10.31 ± 1.35	10.27 ± 1.44	0.744	0.028	0.326
PLT	196.85(140.04,253.66)	198.98(137.87,260.09)	0.682	0.035	−0.409
HGB	140.88(122.20,159.56)	138.09(122.33,153.85)	< 0.05	0.176	2.053
AGE	63.96(49.44,78.48)	60.70(45.73,75.67)	< 0.05	0.218	2.585
SEX	1.31 ± 0.46	1.57 ± 0.50	< 0.05	0.525	−6.234

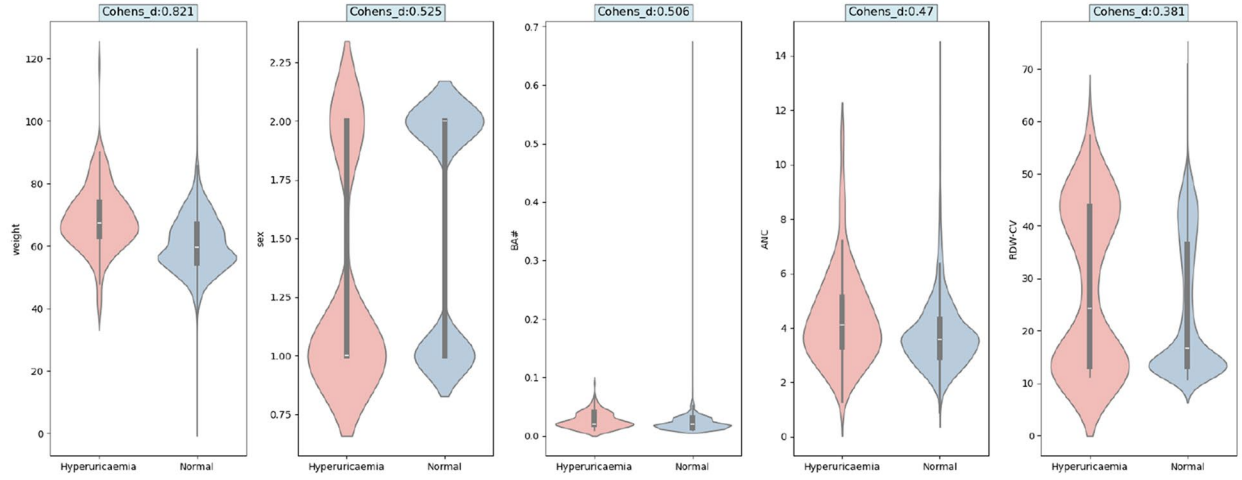

Fig. 2 The top 5 features with significant differences between hyperuricemia and normal samples

RDW-CV, as shown in Fig. 2. Figure 3 shows the correlation of the top 5 features, with the correlation coefficient ranging from 0 to 1, where a value closer to 0 implies a weaker correlation.

However, due to the differences in testing indicators across various hospitals, there are a large number of NULL values in the aforementioned characteristics within the dataset. Since the LR method cannot handle NULL values, it is necessary to process these NULL values. In this paper, multiple imputation is used to address this issue.

Multiple imputation is a statistical technique for handling missing values in a dataset, which mainly consists of three stages: imputation, analysis, and pooling. Firstly, the missing values are predicted and imputed using observed values, generating multiple complete duplicate

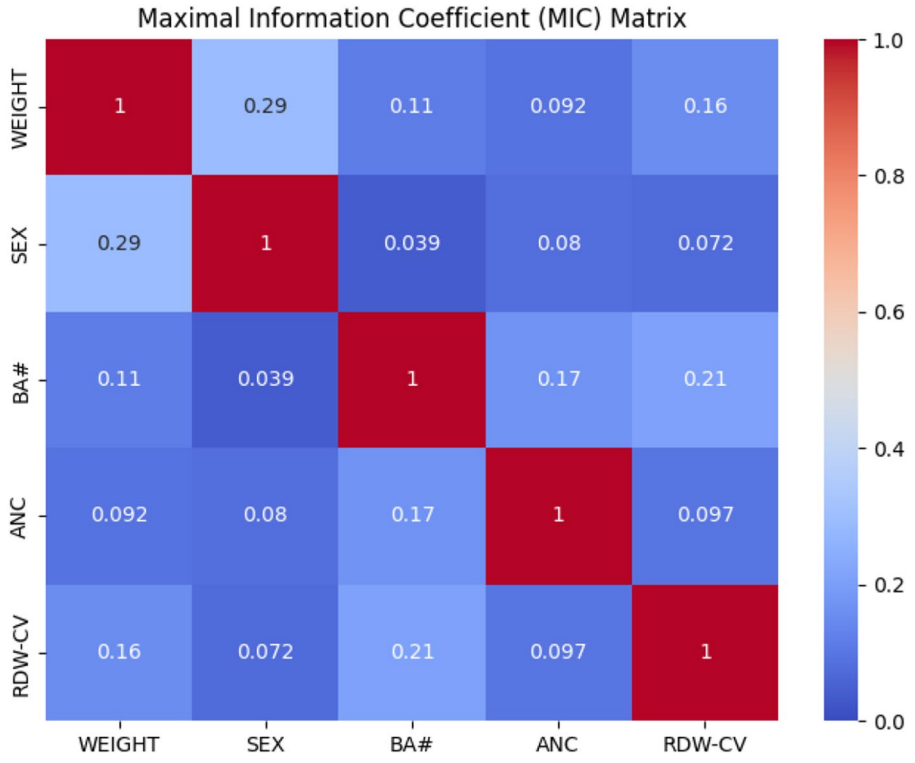

Fig. 3 Heat map of the features

Table 3 Multivariate logistic regression analysis of hyperuricemia

Indicators	Coef	Std_err	Z	P	95%CI
CONST	4.868	6.625	0.735	0.462	[−8.117 17.854]
HGB	−0.003	0.006	−0.416	0.677	[−0.015 0.010]
LYM	−0.133	0.069	−1.934	0.053	[−0.269 0.002]
NEUT%	−0.155	0.074	−2.093	0.036	[−0.299 − 0.010]
ANC	0.477	0.126	3.782	0.000	[0.230 0.724]
MONO#	−1.804	1.144	−1.577	0.115	[−4.047 0.438]
BASO#	−0.454	3.439	−0.132	0.895	[−7.194 6.286]
EOS#	−1.724	1.126	−1.532	0.126	[−3.930 0.482]
RDW-CV	0.021	0.007	3.059	0.002	[0.008 0.035]
AGE	0.037	0.008	4.640	0.000	[0.021 0.052]
SEX	−0.428	0.238	−1.799	0.072	[−0.894 0.038]
WEIGHT	0.031	0.007	4.425	0.000	[0.017 0.045]

datasets. Secondly, each duplicate dataset is analyzed using appropriate statistical methods. Finally, the best imputation method is selected based on model scoring, and the analysis results are pooled to obtain the final statistical inference. After imputation, factors and indicators with statistical significance are assigned values for multivariate logistic regression analysis. The gender variable is assigned with values of male = 1 and female = 2, while all other variables can be substituted with their original values.

Table 3 displays the results of the multivariate logistic regression. It can be observed from the table that LYM, NEUT%, ANC, RDW-CV, AGE, SEX, and WEIGHT are associated with the prevalence of hyperuricemia.

Model construction

Baseline model selection

To find the most suitable risk prediction model for hyperuricemia, this paper selects five machine learning methods as the baseline models, including RF, XGBoost, SVM, LR, and DNN [43, 44].

LR is a widely used linear classification algorithm that converts outputs to probabilities through the sigmoid function, making it highly suitable for binary classification problems. RF is an ensemble learning method that constructs a multitude of decision trees to perform classification or regression tasks. RF mitigates the risk of overfitting by training each decision tree independently on random subsets of samples and features, and it enhances the overall model's accuracy and robustness by aggregating the predictions of these trees [45]. XGBoost is an optimized gradient-boosting decision tree algorithm that improves upon the traditional Gradient Boosting Decision Tree (GBDT) by increasing the training speed and the model's generalization ability. SVM is a powerful linear classifier that uses various kernel functions, such as linear, polynomial, and radial basis functions, to transform the original feature space into a high-dimensional space, and it completes the classification task by finding

the optimal separating hyperplane in this high-dimensional space. DNN consists of an input layer, hidden layers, and an output layer, with each layer having numerous neurons that receive inputs from other neurons. By adjusting the weights, the influence of inputs on the neurons is altered. Neural networks approximate complex functions through multiple nonlinear hidden layers. DNN calculates the error between the output layer and the true labels and propagates the error back to each layer's neurons, updating the neuron weights and bias terms to minimize the prediction error.

Parameter tuning for the baseline models

Parameter tuning is an important way to improve the performance of machine learning models, as it can help the models achieve better generalization capabilities thus significantly enhancing the predictive accuracy [46]. This paper applies grid search and two biological heuristic methods for parameter adjustment of the aforementioned learning models.

Grid search determines the optimal parameters by trying all possible combinations of parameters, which is a simple and effective method for hyperparameter tuning and is widely used in machine learning parameter tuning. Biological heuristic algorithms are a type of optimization algorithm that simulates mechanisms such as natural selection, adaptation, and evolution in biological systems. They possess strong global search capabilities and good adaptability, providing a powerful auxiliary tool for the parameter tuning of machine learning models, especially when dealing with high-dimensional, complex, non-convex, and multi-objective optimization problems. During the training process, this paper uses two types of biological heuristic algorithms including PSO and Genetic Algorithm (GA) to search for the optimal model parameters. Following is a brief introduction of PSO and GA algorithms.

PSO is an evolutionary computation technique to find the optimal solution through collaboration and information sharing among individuals in the group. PSO has been widely applied in function optimization, neural network training, and other application fields. A group of random particles is initialized in the PSO algorithm and searches for the optimal solution independently. The individual extreme value namely pbest is shared with other particles in the swarm, and the best individual extreme is found and used as the current global optimal solution (gbest) for the entire swarm. All particles in the swarm update themselves by tracking the two "extremes" (pbest, gbest).

Equations (2) and (3) show the update method of PSO.

$$v_i = v_i + c_1 \times \text{rand}() \times (pbest_i - x_i) + c_2 \times \text{rand}() \times (gbest_i - x_i) \quad (2)$$

$$x_i = x_i + v_i \quad (3)$$

In Equation (2), v_i represents the current velocity of the particle, c_1 and c_2 are learning factors, rand() is a random number between 0 and 1, and x_i is the current position of the particle. Equation (2) consists of three parts: the memory term, the self-cognition term, and the social cognition term. The memory term comes from the velocity and direction of the last iteration, and the self-cognition term is a vector from the current point to the particle's own best point, indicating the part of the particle's movement that comes from its own experience, the social cognition term is a vector from the current point to the swarm's best point, indicating the collaborative cooperation and knowledge sharing among particles. Particles decide their next movement through their own experience and the best experience among their peers.

GA is another type of biological heuristic algorithm. It achieves population optimization by simulating operations such as selection, crossover, and mutation in the biological evolution process, making it suitable for complex global optimization problems. The core elements of the genetic algorithm include chromosome encoding, fitness function, and genetic operators. Chromosome encoding maps the solution space of a problem onto chromosomes, and the fitness function reflects the quality of individual solutions. Genetic operators include selection operators, crossover operators, and mutation operators. The selection operator selects excellent individuals based on their fitness; the crossover operator generates new individuals through gene exchange; the mutation operator performs small probability mutations on individuals to increase population diversity.

Experimental environment deployment and model performance evaluation

The programming language used in this experiment is Python, and packages such as scikit learn, pytorch and xgboost were utilized. The CPU used is 11th Gen Intel (R) Core (TM) i7-11700 @ 2.50 GHz 2.50 GHz, the GPU is NVIDIA GeForce RTX 3060, and the memory is 32GB.

The selected feature values and processed training data were incorporated into five machine learning algorithms (RF, XGBoost, SVM, LR, and DNN) to construct risk prediction models for hyperuricemia. These models were then applied to the validation data set, and the performance of different models was evaluated using common assessment metrics including AUC, accuracy, recall, precision, and F1 score [47].

Table 4 displays the accuracy scores of 10 folds for the models. Overall, The combination of the xgboost classifier and ADASYN method performs better in the first few folds, while the combination of DNN and ADASYN performs better in the later stages. Table 5 shows the

Table 4 Accuracy of each model for 10 folds

Models	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10
Lr + smote	0.761	0.757	0.779	0.759	0.739	0.764	0.789	0.776	0.773	0.740
lr + adasyn	0.759	0.759	0.745	0.754	0.750	0.746	0.756	0.764	0.760	0.748
Svm + smote	0.916	0.934	0.918	0.900	0.895	0.928	0.936	0.927	0.909	0.922
Svm + adasyn	0.934	0.924	0.912	0.941	0.922	0.932	0.926	0.934	0.926	0.927
Rf + smote	0.927	0.919	0.923	0.921	0.913	0.923	0.942	0.935	0.934	0.912
Rf + adasyn	0.923	0.940	0.935	0.943	0.934	0.942	0.937	0.930	0.924	0.922
Xgboost + smote	0.962	0.974	0.969	0.964	0.951	0.967	0.969	0.964	0.970	0.962
Xgboost + adasyn	0.969	0.974	0.975	0.974	0.970	0.971	0.969	0.966	0.971	0.975
Dnn + smote	0.901	0.967	0.965	0.956	0.968	0.976	0.968	0.986	0.976	0.981
Dnn + adasyn	0.918	0.965	0.964	0.970	0.968	0.979	0.984	0.986	0.991	0.984

Table 5 Performance comparison of five machine learning models with data balancing technology

Models	Auc	Acc	*P*	R	F1
LR + smote	0.855	0.777	0.743	0.846	0.791
LR + adasyn	0.844	0.742	0.678	0.922	0.781
SVM + smote	0.972	0.921	0.921	0.921	0.921
SVM + adasyn	0.975	0.927	0.914	0.943	0.928
RF + smote	0.984	0.934	0.927	0.941	0.935
RF + adasyn	0.982	0.933	0.929	0.937	0.932
DNN + smote	0.958	0.958	0.947	0.969	0.954
DNN + adasyn	0.938	0.938	0.917	0.963	0.960
XGBoost + smote	0.997	0.973	0.975	0.971	0.973
XGBoost + adasyn	0.995	0.972	0.986	0.957	0.971

results of the validation set using the data balancing technology after 10-fold on the metrics mentioned above. The results indicate that the XGBoost model combined with SMOTE is the best one, along with the highest scores in AUC (0.997), accuracy (0.973), Recall (0.971), and f1(0.973). And except for SVM, the others that apply SMOTE achieve better results than ADASYN.

Table 6 shows the results of using different parameter tuning methods (grid search, GA, and PSO) for the XGBoost model, indicating that the PSO algorithm achieves the best results in AUC, ACC, and P indicators. Therefore, in this paper, PSO was selected as the model parameter tuning method. The PSO-based parameter optimization algorithm for machine learning models is shown in Algorithm 2. The initialization P_i of the particle swarm is a randomly selected point, and the objective function and fitness evaluation are both $f(P_i) = -\text{roc\_auc\_score}(y_{\text{test}}, y_{\text{pred}}(P_i))$, with a stopping criterion of reaching the maximum iteration count of 100.

Table 7 Model parameter settings

Model	Parameters
Logistic	max_iter = 1000,C = 100
SVM	kernel = 'rbf', gamma = 0.1, C = 1.0,probability = True
RF	n_estimators = 106, max_depth = 12, max_features = 0.569, min_samples_leaf = 3,min_samples_split = 8
XGBoost	max_depth = 8,learning_rate = 0.2,n_estimators = 171, subsample = 0.9,colsample_bytree = 0.8
DNN	batch_size = 32,epochs = 100,learning_rate = 0.001, optimizer = adam,loss = CrossEntropyLoss

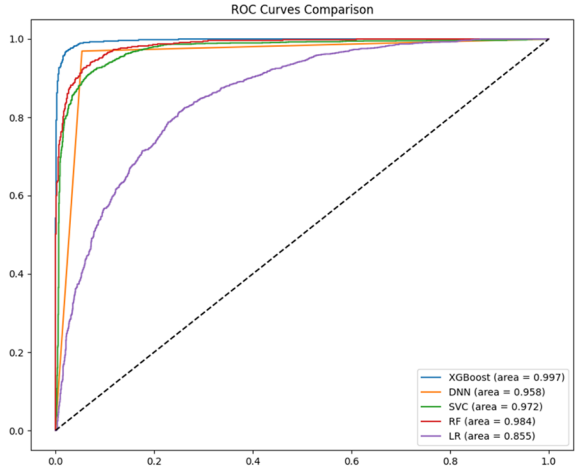

Fig. 4 Comparison of ROC curves for various models

The optimal parameters for each model given by the PSO are shown in Table 7. Figures 4 and 5 display the ROC curves and precision-recall curves.

Table 6 Performance comparison of optimization algorithms applied to XGBoost

optimization	Auc	Acc	P	R	F1	parameters
PSO	0.997	0.973	0.975	0.971	0.973	n_particles = 10,c_1 = 0.5,c_2 = 0.3,w = 0.9,max_iters = 100
GridSearch	0.997	0.972	0.970	0.976	0.976	scoring = f1,cv = 5,n_jobs = −1,verbose = 1
GA	0.996	0.969	0.969	0.968	0.968	n = 10,max_gen = 100,pc = 0.7,pm = 0.2

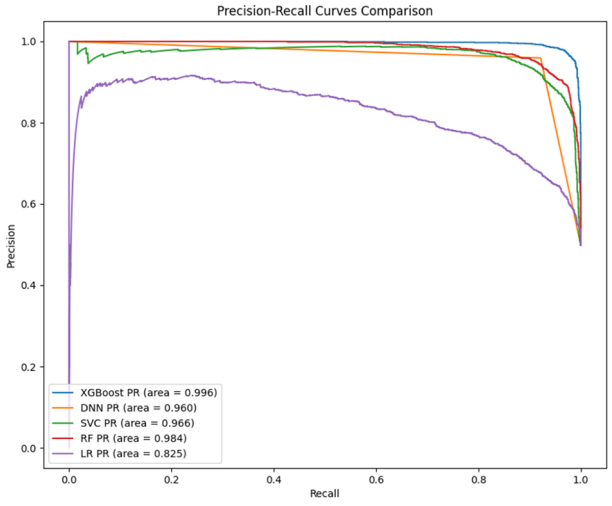

Fig. 5 Comparison of precision-recall curves for various models

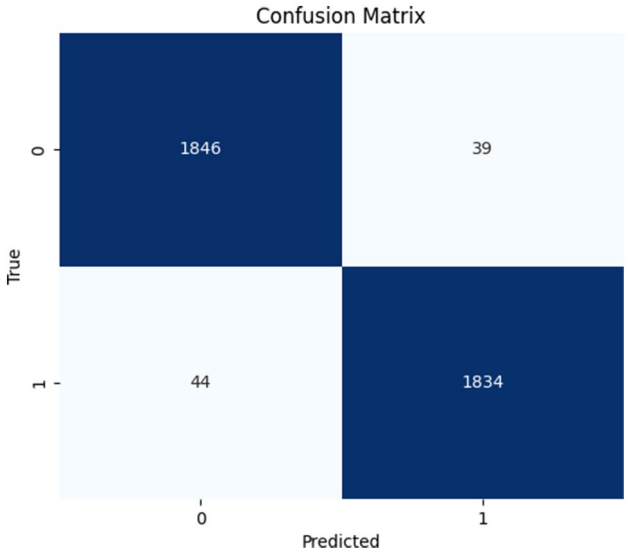

Fig. 6 Confusion matrix of stacking model

Model optimization using model ensemble method

Model ensemble refers to the technique of integrating various machine learning models to enhance the accuracy and stability of predictions. Common model ensemble methods include Bagging, Boosting, Stacking, and Voting [48]. The baseline models mentioned earlier, RF and XGBoost, belong to the Bagging and Boosting methods, respectively. Therefore, this section only applies the other two methods, Stacking and Voting, to improve prediction accuracy.

Voting balances the prediction results of multiple models through probability information. Combining the model comparison results from Section 3.3, the LR model performed the worst, while the XGBoost model performed the best. Thus, weights of 0.15, 0.25, and 0.2 were assigned to the LR, XGBoost, and the other three models, respectively. Stacking combines multiple basic learners and uses a meta-learner to optimize the final prediction, thus fully utilizing the diversity of different models and improving overall prediction performance. In this section, the aforementioned five models are used as basic learners, and logistic regression is chosen as the meta-learner.

Table 8 shows the performance of the Voting and Stacking ensemble models. After applying the model ensemble method, although the AUC decreased, the other four indicators improved to varying degrees. Among them, Stacking achieved the highest ACC, P, and F1 values, while Voting achieved the highest R-value.

Algorithm 2 PSO-based Optimization Algorithm for Machine Learning Models

Input: Dataset T after SOMTE, Optimized model M, hyperparameter ranges R, number of particles *n_particles*
Output: Model *M_best* using optimal parameters
for $i = 1$ **to** *n_particles* **do**
    Initialize particles' positions X_i within the hyperparameter ranges R
    Initialize particles' velocities V_i to zero
end for
while not converged **do**
    for each particle **do**
        Evaluate the fitness of each particle using model M and dataset T
        Update personal best position for each particle
    end for
    Update global best position among all particles
    for $i = 1$ **to** *n_particles* **do**
        Update velocity V_i based on personal best and global best positions
        Update position X_i using velocity V_i
    end for
end while
Apply optimal hyperparameters to the model M
return *M_best*

Table 8 Performance of the voting and stacking ensemble models

Ensemble models	AUC (95%CI)	Accuracy (95%CI)	Precision (95%CI)	Recall (95%CI)	F1 (95%CI)
Stacking	0.996(0.995,0.998)	0.978(0.973,0.982)	0.980(0.972,0.985)	0.976(0.969,0.982)	0.978(0.972,0.982)
Voting	0.996(0.994,0.997)	0.973(0.968,0.978)	0.961(0.952,0.969)	0.987(0.982,0.992)	0.974(0.968,0.978)

Table 9 Comparison with state-of-the-art methods

Papers	Feature optimization	Number of features	Classifier	Accuracy	AUC	XAI
Hou et al. [4]	Lr analysis	15	SVM	0.819	0.875	No
Zheng et al. [27]	LASSO	6	XGBoost	0.881	0.733	No
Zeng et al. [28]	Lr analysis	14	ANN	0.800	0.814	No
Ma [9]	Lr analysis	10	Bayesian	-	0.740	No
Yang [11]	LASSO	4	Lr	0.726	0.813	No
Chen et al. [49]	-	14	XGBoost	0.730	0.820	No
Gao et al. [5]	-	21	Rf	-	0.739(male) 0.818(Female)	No
Proposed model	Lr analysis	7	Stacking	0.978	0.978	Yes

A confusion matrix helps visualize the output of a classifier. It depicts a table of accurate detections and misdetections of an ML model. Figure 6 shows the confusion matrix for our best model.

The training and testing data used in this paper come from real data from multiple hospitals in two regions of Zhejiang Province. However, due to the limitations of the dataset, there may be concerns about the generalization and universality of the model. To address this, we used the Mostly (http://www.mostly.ai) synthetic data genera tion technique to generate an equal amount of data for testing based on the data characteristics of the validation set. After testing, the final accuracy of the model reached 0.820, with an AUC value of 0.879, confirming that the model still performs well on the generated data.

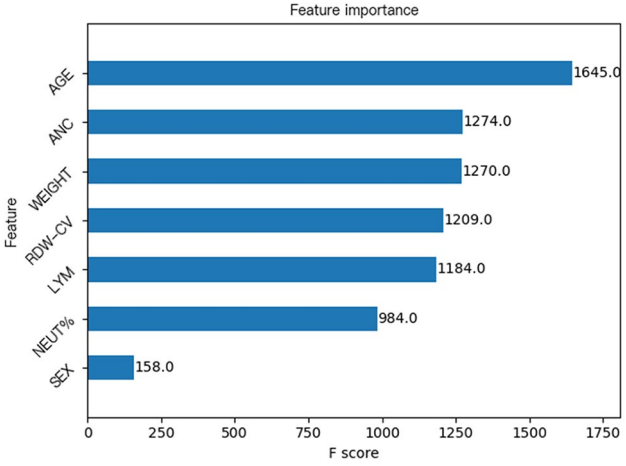

Fig. 7 Feature importance ranking

Performance comparison with state-of-the-art models

To demonstrate the advantage of the fusion model proposed in this paper, we also compare its performance with several other state-of-the-art models used for the risk prediction of hyperuricemia.

The comparison models selected in this paper are from references [4, 9, 11, 27, 28, 49] and [5]. Hou et al. [4] constructed a hyperuricemia risk prediction model based on the SVM method and included fifteen variables in the model for prediction through the LR technique. Zheng et al. [27] used the LASSO method to incorporate six variables into XGBoost to construct a hyperuricemia risk prediction model. Zeng et al. [28] selected fourteen variables through the LR analysis and then generated a hyperuricemia risk prediction model based on the ANN method. Ma [9] also selected ten variables through the LR and subsequently constructed a prediction model based on the Bayesian method. Yang's model [11] hyperuricemia risk prediction model, based on LR technology, included four feature variables through LASSO analysis. Chen et al. [49] primarily constructed was based on the XGBoost method, which included 14 variables. Gao et al. [5] designed a hyperuricemia risk prediction model based on the RF method, incorporating 21 variables.

Table 9 shows the comparison results on methods and performance. It can be seen that compared to previous studies, our model achieved better results, with an improvement of over 11% in predictive accuracy.

Model interpretability

Medical issues are directly related to patient's health and life safety, making it a field that demands high accuracy and interpretability. Medical personnel need to understand and interpret the diagnostic results produced by AI models to ensure the accuracy and reliability of the diagnostic process.

To enhance the interpretability, this paper uses the XGBoost model to obtain the feature importance for the risk of hyperuricemia. The feature importance scores in XGBoost can be calculated using three methods: (1) Weight, indicating the number of times a feature is used as a splitting feature in decision trees; (2) Gain, representing the contribution of a feature to the performance gain of the model during the split; and (3) Coverage, indicating the proportion of data samples covered by the feature when it is used for splitting. This paper uses the weight-based feature importance ranking, and Fig. 7 plots the results. The features in the figure are sorted in descending order of importance. Among all the features, age is the most important, significantly impacting the prediction results, while gender has the smallest impact.

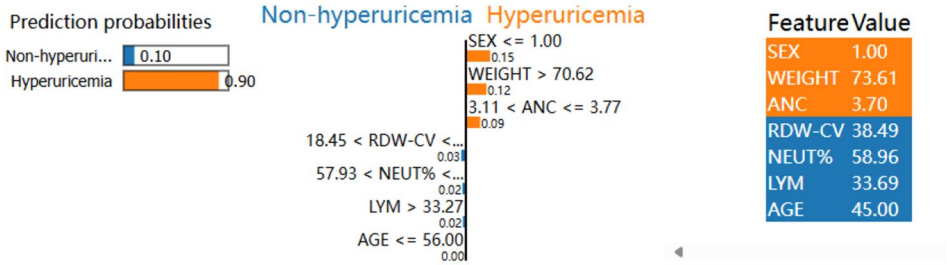

Fig. 8 LIME plot

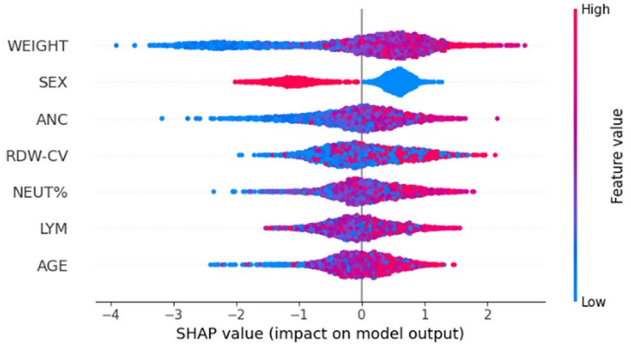

Fig. 9 SHAP feature density scatter plot

Using XGBoost to get feature importance is quick and simple, thus it is suitable for a preliminary assessment of feature importance. However, it has some shortcomings. For example, when features are highly correlated, XGBoost may overestimate or underestimate the importance of some specific features. This is because information may be double-counted, leading to inaccuracies in the importance of features with multicollinearity. Additionally, XGBoost feature evaluation often considers the impact of features in isolation, without assessing the interaction effects between features, neglecting feature interactions. Therefore, a more precise and fine-grained explanation method is needed when decisions affect individual disease diagnosis.

We implemented several XAI methods including SHAP and LIME to reach the goal. SHAP is based on the Shapley value from game theory and is used to measure the contribution of each feature to the model's prediction, thereby helping people understand how the model makes decisions [50]. LIME is an algorithm that interprets classifier or regressor predictions by locally approximating them with interpretable models. It analyzes the contribution of features by adjusting feature values and observing their impact on the output. The output of LIME is a set of explanations representing the contribution of each feature to the prediction of a single sample, which aids in understanding the model's predictions.

Figure 9 is a feature density scatter plot generated using the SHAP method. It indicates that WEIGHT, ANC, RDW-CV, NEUT%, LYM, and AGE significantly impact the model output. By examining the SEX row, we found that the SHAP values for males (blue points) slightly lean towards the positive side, while the SHAP values for females (red points) slightly lean towards the negative side. This suggests that in the model's predictions, males are slightly more likely to be predicted as patients, whereas females are slightly more likely to be predicted as healthy Individuals. Figure 8 illustrates the feature importance of our best model for a specific instance based on the LIME method. The horizontal bars represent the contribution of each feature to the prediction; the length of the bars indicates the weight of the feature in the model, and the color represents the direction of the impact (orange for positive, blue for negative). The plot also includes the actual predicted value, the model's predicted value, and the difference between them.

In summary, Fig. 7 provides a preliminary assessment of feature importance. Figure 8 offers detailed prediction information for a specific sample, including feature values and their specific contributions to the prediction outcomes. Figure 9, on the other hand, provides a global perspective, demonstrating the impact of features from all samples in the test set on the model. It can be observed from Figs. 8 and 9 that for both individual and global samples, SEX and WEIGHT significantly contribute to the model's predictions, ANC makes a certain contribution, and AGE contributes less. The above conclusion is consistent with humans suffering from hyperuricemia. Medically, it is believed that estrogen has an excretory effect on uric acid, which is why women are less likely to develop hyperuricemia compared to men. Obesity is usually related to increased uric acid synthesis and reduced renal excretion. However, it is noteworthy that in Fig. 7, the contribution of SEX is the smallest, whereas the contribution of AGE is the largest. This paper believes that the root cause of inconsistency lies in the different calculation methods and concerns of interpretation methods. The feature importance of xgboost mainly reflects the role of features in the process of model construction and is affected by the three calculation methods mentioned above. SHAP and LIME provide explanations from the perspective of contribution degree, locality, or globality of model output, especially considering feature

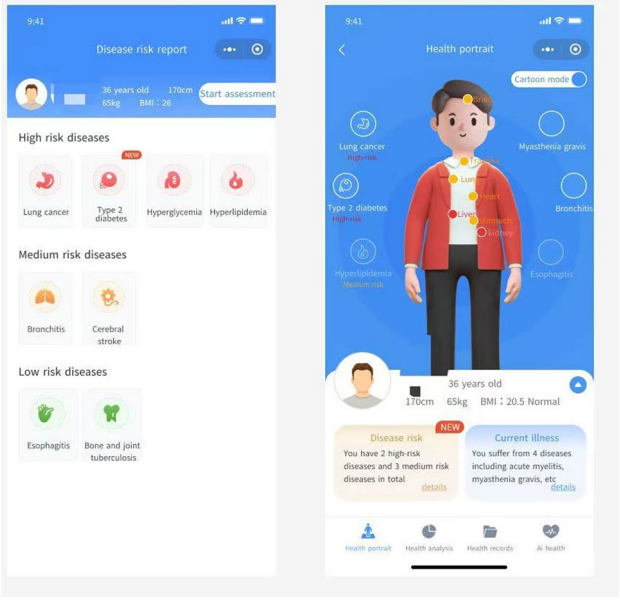

Fig. 10 Health portrait home page and disease risk assessment page

interaction and nonlinear structure of the model, which can better reveal the true relationship between features and prediction results. The common point is that both ANC and WEIGHT have a significant impact.ke

Disease risk prediction platform

This paper has conducted a predictive model for hyperuricemia, and we hope to apply it in practical scenarios, allowing more patients to receive timely warnings and interventions before the onset or rapid progression of the disease. To this end, we have developed a Health Portrait Platform, which has the core functionality of incorporating disease risk prediction models for analysis based on input features, and the main features such as storage and management of residents' electronic health records, disease risk assessment and prediction, early cancer screening, online diagnosis and treatment, and a health mall.

Till now, the platform has covered data for 650,000 permanent residents of Lin'an, Hangzhou, China with about 10,000 users, and the number of visitors has reached approximately 100,000. Figure 10 displays the homepage and disease risk assessment page of the Health Portrait platform's user interface. The homepage shows the user's current diseases and potential risk diseases. The disease risk assessment page will provide a detailed display of the low-, medium-, and high-risk disease lists predicted by the model.

Conclusion and future work

This paper constructs electronic health records based on the physical examination and patient treatment data from the hospitals in two districts of Hangzhou, China, and builds a hyperuricemia risk prediction model using machine learning algorithms. Experimental validation results indicate that this prediction model has good predictive accuracy, providing a reference for the early identification of risk factors for hyperuricemia. The innovations of this model include: 1) Addressing dataset deficiencies through data processing and interpolation; 2) Selecting the optimal model through preliminary comparative analysis of different AI algorithms, while using the PSO algorithm to fine-tune the parameters of the machine learning models, thereby improving accuracy and generalizability; 3) Modeling and analyzing based solely on routine blood test indicators, breaking the limitation of traditional hyperuricemia prediction that requires biochemical data, reducing screening costs, and facilitating large-scale implementation; 4) Developing a user-health profiling platform that visualizes health analysis results in an interactive chart format, enabling chronic disease screening for regional populations and individual residents.

Due to data privacy protection, this paper currently only has access to authorized physical examination and medical data from hospitals in two regions of Zhejiang, China. The limited data may affect the universality of the proposed prediction model. In addition, the data has a high imbalance rate, the sample size of hyperuricemia cases is too small, and the data synthesized after the application of the SMOTE method may distort the actual distribution. Moreover, using only routine blood tests for hyperuricemia risk assessment may have certain limitations in practical applications.

Currently, we have embedded the model into a disease risk prediction platform that can obtain relevant data filled in by users and conduct accuracy assessments in a real-time manner. In the future, we will actively seek more relevant or public data to adjust and optimize the model, enhancing its universality and generalization capabilities. Secondly, we will continue to refine the processes and parameters of the prediction model to improve its accuracy and performance. Lastly, we will keep enhancing the functionality of the disease risk prediction platform to achieve broader applications.

Acknowledgments
The authors would like to thank the hard work of the editors and reviewers.

Author contributions
Min Fang, Chengjie Pan, Wenjuan Li, and Ben Wang wrote the main manuscript text, Xiaoyi Yu, Huajian Zhou, and Zhenying Xu developed the disease risk prediction platform, and Genyuan Yang collected the data and prepared the figures in this paper. All authors reviewed the manuscript.

Funding
This research was funded in part by the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant LHZSZ24F020001, in part by the National Natural Science Foundation of China under Grant 61702151, in part by the Noncommunicable Chronic Diseases-National Science and Technology Major Project under Grant 2023ZD0509800, and in part by the

Annual Planning Project for 2024 of Engineering Research Center of Mobile Health Management System, Ministry of Education.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Ethics approval for this study was approved by the Ethics Committee of Hangzhou Normal University, which belongs to Hangzhou Normal University. All participants were provided with information regarding the study and gave their written informed consent before participation. This study was conducted in compliance with the Declaration of Helsinki and all applicable ethical guidelines.

Consent for publication
Informed consent was obtained from all subjects involved in the study.

Competing interests
The authors declare no competing interests.

Received: 5 November 2024 / Accepted: 28 February 2025
Published online: 14 March 2025

References

1. Chen-Xu M, Yokose C, Rai S, et al. Contemporary prevalence of gout and hyperuricemia in the united states and decadal trends: the national health and nutrition examination survey, 2007–2016. Arthritis Rheumatol 2019;71:991–99.
2. Fang NY, Lv LW, Lv XX, et al. Chinese expert consensus on the diagnosis and treatment of hyperuricemia-related diseases (2023 edition). Chin J Pract Internal Med. 2023;43(06):461–80. https://doi.org/10.19538/j.nk2023060106.
3. Conen D, Wietlisbach V, Bovet P, et al. Prevalence of hyperuricemia and relation of serum uric acid with cardiovascular risk factors in a developing country. BMC Public Health. 2004;4(9). https://doi.org/10.1186/1471-2458-4-9.
4. Hou K, Shi Z, Ge X, Song X, Yu C, Su Z, Wang S, Zhang J. Study on risk factor analysis and model prediction of hyperuricemia in different populations. Front Nutr. 2024; 11. https://doi.org/10.3389/fnut.2024.1417209.
5. Gao Y, Jia S, Li D, Huang C, Meng Z, Wang Y, Yu M, Xu T, Liu M, Sun J, Jia Q, Zhang Q, Gao Y, Song K, Wang X, Fan Y. Prediction model of random forest for the risk of hyperuricemia in a chinese basic health checkup test. 41(4):20203859. https://doi.org/10.1042/BSR20203859. Accessed 19 Dec 2024.
6. Li J. Construction of a risk prediction model for adult hyperuricemia based on a nomogram. PhD thesis, Hubei University of Medicine. 2023.
7. Liu ZY, Wang JM, Wei YP, et al. Relationship between sleep duration and hyperuricemia in community residents. Chinese General Pract 2022;25:1681–86.
8. Zhu B, Yang L, Wu M, Wu Q, Liu K, Li Y, Guo W, Zhao Z. Prediction of hyperuricemia in people taking low-dose aspirin using a machine learning algorithm: a cross-sectional study of the national health and nutri- tion examination survey. Front Pharmacol. 2024;14:1276149. https://doi.org/10.3389/fphar.2023.1276149.
9. Ma Y. Analysis of influencing factors and construction of a risk prediction model for hyperuricemia among employees in a comprehensive hospital in shanghai. PhD thesis, Nanchang University. 2023. https://doi.org/10.27232/d.cnki.gnchu.2023.000800.
10. Peng C. Research on the risk prediction model of hyperuricemia. PhD thesis, Southern Medical University. 2023. https://doi.org/10.27003/d.cnki.gojyu.2023.000508.
11. Yang K. Construction of a risk prediction model for hyperuricemia among construction workers. PhD thesis, Hubei University of Medicine. 2023. https://doi.org/10.27913/d.cnki.ghyby.2023.000076.
12. Yu H, Zhang J, Liu F, et al. Hyperuricemia risk prediction model based on longitudinal health examination data. Mod Preventive Med 2021;48:4408–12.
13. Wang Y, Zeng Y, Zhang X, et al. Daytime napping duration is positively associated with risk of hyperuricemia in a chinese population. 2023.
14. Shi JC, Chen XH, Yang Q, et al. A simple prediction model of hyperuricemia for use in a rural setting. Sci Rep. 2021;11:23300. https://doi.org/10.1038/s41598-021-02716-y.
15. Zeng L, Ma P, Li Z, Liang S, Wu C, Hong C, Li Y, Cui H, Li R, Wang J, He J, Li W, Xiao L, Liu L. Multimodal machine learning-based marker enables early detection and prognosis prediction for hyperuricemia. Adv Sci (Weinh). 2024;2024:2404047. https://doi.org/10.1002/advs.202404047.
16. Geng YH, Zhang Z, Zhang JJ, et al. Established the first clinical prediction model regarding the risk of hyperuricemia in adult iga nephropathy. Int Urol Nephrol. 2023;55:1787–97. https://doi.org/10.1007/s11255-023-03498-0.
17. Chen C, Sun L, Tao Y, Xu C. Predictive value and clinical utility of the risk model containing routine blood and coagulation indicators for sepsis secondary to malignant hematological diseases. Hainan Med. 2024;35(14):2031–36.
18. Hossain MM, Hasan M, Rahim A, Rahman MM, Yousuf MA, Al-Ashhab S, Akhdar HF, Alyami SA, Azad A, Moni MA. Particle swarm opti- mized fuzzy CNN with quantitative feature fusion for ultrasound image quality identification 10.
19. Islam H, Iqbal MS, Hossain MM. Blood pressure abnormality detection and interpretation utilizing explainable artificial intelligence, 2667102624000676. https://doi.org/10.1016/j.imed.2024.09.005. Accessed 13 Dec 2024.
20. Cao J, Wang C, Zhang G, Ji X, Liu Y, Sun X, Yuan Z, Jiang Z, Xue F. Incidence and simple prediction model of hyperuricemia for urban han chinese adults: a prospective cohort study. Int J Environm Res Public Health. 2017;14:67. https://doi.org/10.3390/ijerph14010067.
21. Yaqoob A, Verma NK, Aziz RM. Improving breast cancer classification with mrmr+ ss0+ wsvm: a hybrid approach. Multimedia Tools Appl. 2024;1–26.
22. Yaqoob A, Verma NK, Aziz RM. Metaheuristic algorithms and their applications in different fields: a comprehensive review. Metaheuristics Machine Learning: Algorithms Appl. 2024;1–35.
23. Wang TJ, Massaro JM, Levy D, et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the framingham heart study. JAMA. 2003;290(8):1049–56. https://doi.org/10.1001/jama.290.8.1049.
24. Khosla A, Cao Y, Lin CC, Chiu HK, Hu J, Lee H. An integrated machine learning approach to stroke prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 183–92. 2010. https://doi.org/10.1145/1835804.1835831.
25. Kun L, Chunmei C, Rui F, et al. Detection of diabetic patients in people with normal fasting glucose using machine learning. BMC Med 2023;21:342–342.
26. Lee S, Choe EK, Park B. Exploration of machine learning for hyperuricemia prediction models based on basic health checkup tests. 8(2):172. https://doi.org/10.3390/jcm8020172. Accessed 11 Jan 2025.
27. Zheng Z, Si Z, Wang X, Meng R, Wang H, Zhao Z, Lu H, Wang H, Zheng Y, Hu J, He R, Chen Y, Yang Y, Li X, Xue L, Sun J, Wu J. Risk prediction for the development of hyperuricemia: model development using an occupational health examination dataset. 20(4):3411. https://doi.org/10.3390/ijerph20043411. Accessed 11 Dec 2024.
28. Zeng J, Zhang J, Li Z, Li T, Li G. Prediction model of artificial neural network for the risk of hyperuricemia incorporating dietary risk factors in a chi- nese adult study. Food Nutr Res. 2020;64:3712. https://doi.org/10.29219/fnr.v64.3712.
29. Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: a review. 1808;11(12). https://doi.org/10.3390/healthcare11121808. Accessed 11 Jan 2025.
30. Zhang Y, Zhang L, Lv H, Zhang G. Ensemble machine learning prediction of hyperuricemia based on a prospective health checkup population. 15:1357404. https://doi.org/10.3389/fphys.2024.1357404. Accessed 11 Jan 2025.
31. Su XQ, Shi XL, Zhou MH, Xu YQ. Application effect of comprehensive health education in patients with hyperuricemia in the community. J Lab Med Clin Med. 2022;19(9):1249–51. https://doi.org/10.3969/j.issn.1672-9455.2022.09.026.
32. Shirasawa T, Ochiai H, Yoshimoto T, et al. Cross-sectional study of asso- ciations between normal body weight with central obesity and hyperuricemia in japan. BMC Endocr Disord 2020;20:2.
33. Xiao N, Li L, Guo X, et al. Construction and validation of a risk prediction model for hyperuricemia in children and adolescents. Nurs Res 2024;38:1908–13.
34. Zheng Z. Research on the risk prediction of hyperuricemia among steel workers. PhD thesis, North China University of Technology. 2023. https://doi.org/10.27108/d.cnki.ghelu.2023.000301.

35. Yang K, Zhu G, Liu X, Zhou S, Wang C, Liu B. Establishment of a lasso-logistic regression prediction model for hyperuricemia. Med Rev. 2023;29(18):3708–14.
36. Song X, Jiang Z, Ge W, et al. Construction of a risk prediction model for hyperuricemia. Chinese Primary Health Care 2022;36:51–54.
37. Wang H. Comparison of the application effects of different machine learning models in predicting hyperuricemia in the natural population of the northeast region. PhD thesis, China Medical University. 2022. https://doi.org/10.27652/d.cnki.gzyku.2022.000479.
38. Hallowell N, Badger S, Sauerbrei A, Nellåker C, Kerasidou A. "i don't think people are ready to trust these algorithms at face value": trust and the use of machine learning algorithms in the diagnosis of rare disease. BMC Med Ethics. 2022;23(1):112.
39. Salloch S, Heyen NB. The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. BMC Med Ethics. 2021;22(1):1–9.
40. Borghi C, Verardi FM, Pareo I, Bentivenga C, Cicero AF. Hyperuricemia and cardiovascular disease risk. Expert Rev Cardiovasc Ther. 2014;12(10):1219–25. https://doi.org/10.1586/14779072.2014.957675.
41. Yaqoob A, Verma NK, Aziz RM, Shah MA. Rna-seq analysis for breast cancer detection: a study on paired tissue samples using hybrid optimization and deep learning techniques. J Cancer Res Clin Oncol. 2024;150(10):455.
42. Yaqoob A, Verma NK, Aziz RM, Shah MA. Optimizing cancer classification: a hybrid rdo-xgboost approach for feature selection and predictive insights. Cancer Immunol Immunother. 2024;73(12):261.
43. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
44. Hamid HA, Nayan NA, Suboh MZ, Aminuddin A: Second derivatives of photoplethysmogram for hyperuricemia classification using artificial neural network. In: 2020 IEEE-EMBS Conference on Biomedical Engineering and Sci- ences (IECBES), IEEE, pp. 494–98. 2021. https://doi.org/10.1109/IECBES48179.2021.9398786.
45. Janitza S, Strobl C, Boulesteix A-L. An auc-based permutation variable importance measure for random forests. BMC Bioinf. 2013;14:119.
46. Aziz RM, Hussain A, Sharma P. Cognizable crime rate prediction and analysis under indian penal code using deep learning with novel optimization approach. Multimedia Tools Appl. 2024;83(8):22663–700.
47. Liang B, Huang Z, Lai Y, et al. Comparison of the application effects of the random forest model and logistic regression model in predicting hyperuricemia. Guangxi Med J. 2020;42(6):729–33. https://doi.org/10.11675/j.issn.0253-4304.2020.06.17.
48. Karadeniz T, Tokdemir G, Maras H. Ensemble methods for heart disease prediction. New Generation Computing. 2021;39:569–81. https://doi.org/10.1007/s00092-021-06257-5.
49. Chen S, Han W, Kong L, Li Q, Yu C, Zhang J, He H. The development and validation of a non-invasive prediction model of hyperuricemia based on mod- ifiable risk factors: baseline findings of a health examination population cohort. Food Funct. 2023;14:6073–82. https://doi.org/10.1039/D3FO01363D.
50. Afreen S, Bhurjee AK, Aziz RM. Cancer classification using rna sequencing gene expression data based on game shapley local search embedded binary social ski-driver optimization algorithms. Microchem J. 2024;205:111280.

Publisher's Note

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