



## OPEN ACCESS

## EDITED BY

Liang Zhao,  
Dalian University of Technology, China

## REVIEWED BY

Elizaveta Savchenko,  
Ariel University, Israel  
Salvatore Annunziata,  
Fondazione Policlinico Universitario A.  
Gemelli IRCCS, Italy

## \*CORRESPONDENCE

Kaiwen Hou  
✉ hkfcc@126.com

RECEIVED 28 April 2024

ACCEPTED 22 October 2024

PUBLISHED 04 March 2025

## CITATION

Tian F, Lin Y, Wang L, Fang F and Hou K (2025)  
Construction of a risk screening and  
visualization system for pulmonary nodule in  
physical examination population based on  
feature self-recognition machine learning  
model.  
*Front. Med.* 11:1424750.  
doi: 10.3389/fmed.2024.1424750

## COPYRIGHT

© 2025 Tian, Lin, Wang, Fang and Hou. This is  
an open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Construction of a risk screening and visualization system for pulmonary nodule in physical examination population based on feature self-recognition machine learning model

Fang Tian<sup>1</sup>, Yongchun Lin<sup>1</sup>, Liangjiao Wang<sup>1</sup>, Fei Fang<sup>2</sup> and Kaiwen Hou<sup>1\*</sup>

<sup>1</sup>Department of Outpatient, Western Theater Command General Hospital of PLA, Chengdu, Sichuan, China, <sup>2</sup>Department of Emergency, Tibet Command General Hospital of PLA, Lhasa, China

**Objective:** To assess the effectiveness of a feature self-recognition machine learning model in screening for pulmonary nodule risk in a physical examination population and to evaluate the constructed visualization system.

**Methods:** We analyzed data from 4,861 individuals who underwent chest CT exams during their physical examinations at the Western Theater General Hospital of the People's Liberation Army from January 2023 to November 2023. Among them, 1,168 had positive CT reports for pulmonary nodules, while 3,693 had negative findings. We developed a machine learning model using the XGBoost algorithm and employed an improved sooty tern optimization algorithm (ISTOA) for feature selection. The significance of the selected features was evaluated through univariate analysis and multivariable logistic stepwise regression analysis. A visualization system was created to estimate the risk of developing pulmonary nodules.

**Results:** Multivariable analysis identified older age, smoking or passive smoking, high psychological stress within the past year, occupational exposure (e.g., air pollution at the workplace), presence of chronic lung diseases, and elevated carcinoembryonic antigen levels as significant risk factors for pulmonary nodules. The feature self-recognition machine learning model further highlighted age, smoking or passive smoking, high psychological stress, occupational exposure, chronic lung diseases, family history of lung cancer, decreased albumin levels, and elevated carcinoembryonic antigen as key predictors for early pulmonary nodule risk, demonstrating superior performance.

**Conclusion:** The feature self-recognition machine learning model effectively aids in the early prediction and clinical identification of pulmonary nodule risk, facilitating timely intervention and improving patient prognosis.

## KEYWORDS

machine learning, pulmonary nodules, risk screening, visualization system, algorithm

## 1 Introduction

The widespread implementation of lung cancer screening programs has markedly increased the detection rates of pulmonary nodules. These nodules, characterized as focal, round-shaped, solid or subsolid lung opacities not exceeding 3 cm in diameter on imaging, can evolve into malignant tumors if not diagnosed and managed promptly. This progression significantly deteriorates the quality of life for affected individuals (1, 2). Lung cancer remains the most prevalent and deadliest of all malignant tumors, with most patients presenting at advanced stages, resulting in low five-year survival rates and poor prognoses (3, 4). Consequently, the effective management of pulmonary nodules is crucial in the prevention and control of lung tumors.

The clinical manifestations of pulmonary nodules are non-specific, complicating the diagnostic process and increasing the likelihood of misdiagnosis. Traditionally, the assessment of these nodules for benign or malignant characteristics involves the analysis of chest CT images or the employment of invasive techniques such as surgery or biopsy to obtain a definitive lesion characterization (5, 6). However, recent advancements in artificial intelligence (AI) have facilitated the extraction of feature information and the development of predictive models. These innovations are proving instrumental in aiding physicians to diagnose suspicious pulmonary nodules non-invasively. Such technological progress not only enhances the potential for early disease detection and prognosis but also significantly improves the diagnostic accuracy of pulmonary nodules (7, 8). For instance, studies have demonstrated that deep learning models are capable of learning subtle image features from complex imaging data, features that are often elusive to traditional methods (2, 9). These advancements have not only improved the accuracy in distinguishing benign from malignant pulmonary nodules but have also shortened the diagnostic process, providing quicker decision support for patient treatment (10, 11). Furthermore, recent research has explored how improvements in algorithms and model structures can enhance the generalizability and interpretability of diagnostic systems, making their application in clinical practice more widespread and effective (12). These findings not only confirm the potential of artificial intelligence technology in non-invasive diagnostics but also highlight future research directions, specifically how to better integrate these advanced technologies into routine clinical diagnostic processes to improve early disease detection and treatment outcomes (13). The aim of this study is to analyze the value of pulmonary nodules risk screening in physical examination population and the effect of visualization system construction based on feature self-recognition machine learning mode.

## 2 Methods and materials

### 2.1 Study population

A total of 4,861 individuals who underwent chest CT examinations as part of their physical examinations at the Western Theater General Hospital of the People's Liberation Army from January 2023 to November 2023 were included in this study, access to the study data began on January 1, 2024. Among them, 1,168 patients had positive CT reports for pulmonary nodules, while 3,693 patients had negative findings. Inclusion criteria were as follows: (1) Normal mental status, clear cognition, and able to cooperate with inquiries; (2) Complete

clinical data. Exclusion criteria were as follows: (1) Presence of severe diseases such as cardiovascular, liver, or kidney disorders; (2) History of previous tumors; (3) Pregnant or breastfeeding women.

This study was conducted retrospectively, informed consent was waived, and this study was approved by the Ethics Committee of the Western Theater General Hospital of the People's Liberation Army (Approval No.: 2022ky105-3). All data were anonymized.

### 2.2 Data collection

General information and laboratory test results of the study participants were obtained from the Health Management System of the Western Theater General Hospital of the People's Liberation Army, comprising a total of 33 features. The general information included gender, age, smoking history, alcohol consumption history, place of residence, education level, chronic lung diseases, family history of lung cancer, regular exercise, body mass index, presence of high psychological stress within the past year, and presence of depressive symptoms within the past year. The laboratory test results included carcinoembryonic antigen, thyroid-stimulating hormone, white blood cell count, lymphocyte count, platelet count, hemoglobin, eosinophil count, basophil count, albumin, globulin, albumin-globulin ratio, alanine aminotransferase, aspartate aminotransferase, indirect bilirubin, high-density lipoprotein, low-density lipoprotein, triglycerides, fasting blood glucose, creatinine, blood urea nitrogen, and uric acid.

### 2.3 Feature self-recognition machine learning model

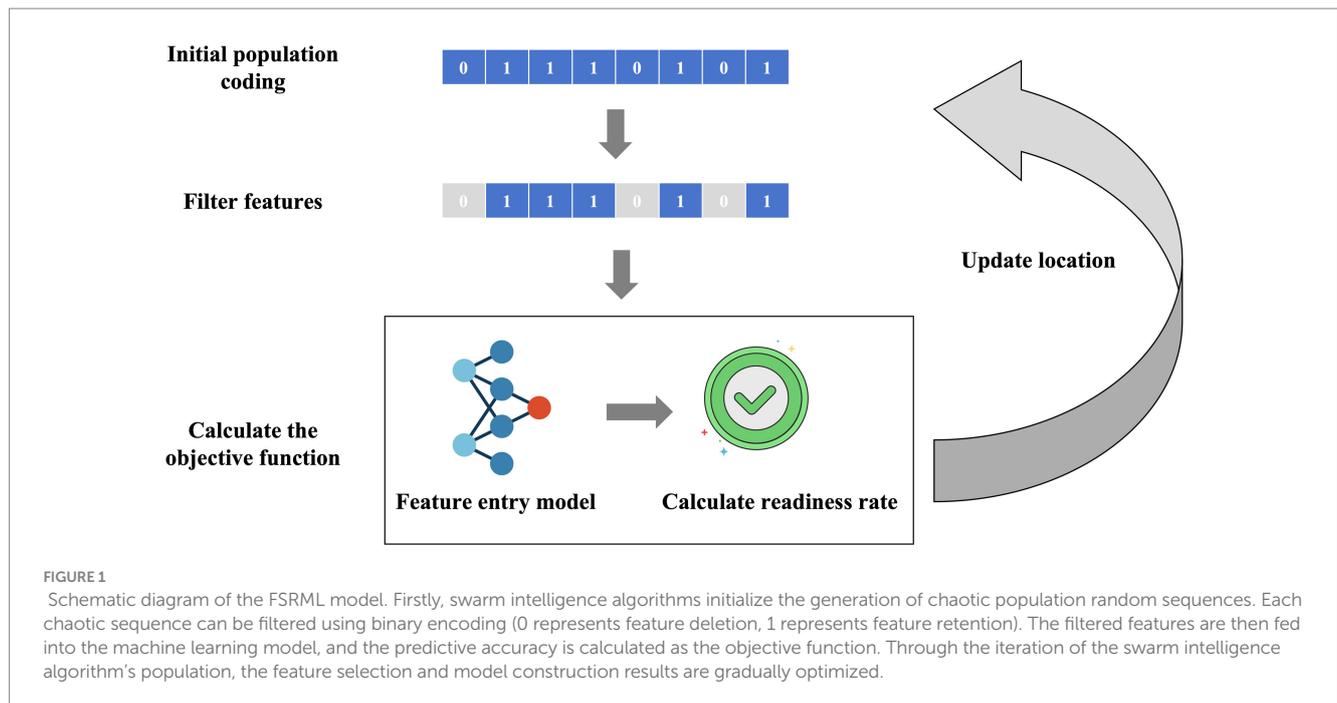
In this study, we proposed a feature self-recognition machine learning model (FSRML) that does not require preliminary feature selection before running. All 33 features studied were included in the model training. The FSRML utilizes the powerful global optimization capability of swarm intelligence algorithms to automatically perform feature selection. The schematic diagram of the model is shown in Figure 1.

Caption: Firstly, swarm intelligence algorithms initialize the generation of chaotic population random sequences. Each chaotic sequence can be filtered using binary encoding (0 represents feature deletion, 1 represents feature retention). The filtered features are then fed into the machine learning model, and the predictive accuracy is calculated as the objective function. Through the iteration of the swarm intelligence algorithm's population, the feature selection and model construction results are gradually optimized.

To better guide the feature optimization task mentioned above, this study improves upon the sooty tern optimization algorithm (STOA) (14) by incorporating three enhancement strategies: Bernoulli chaotic mapping (15), Cauchy mutation perturbation (16), and longitudinal-lateral crossover mutation (17). These improvements lead to the development of a hybrid chaotic sooty tern optimization algorithm that combines longitudinal-lateral crossover and Cauchy mutation. It is referred to as the improved sooty tern optimization algorithm (ISTOA). The specific improvement strategies are as follows:

(1) Bernoulli chaotic mapping

Swarm intelligence optimization algorithms generally generate populations through randomization. However, when the population size is small, the populations generated by random arrays may lack



sufficient ergodicity, potentially causing the optimization results to fall into local optima. Chaotic sequences, characterized by strong ergodicity, unpredictability, and sensitivity to initial values, are better suited for the task of initializing populations. Bernoulli mapping is a typical example of chaotic mapping, and its expression is as follows:

$$x_{n+1} = \begin{cases} \frac{x_n}{1-\lambda} & 0 < x_n < 1-\lambda \\ \frac{x_n - (1-\lambda)}{\lambda} & 1-\lambda < x_n < 1 \end{cases}$$

In the above expression, we set the value to 0.4. First, a random number  $x_0$  between 0 and 1 is generated. Then, the chaotic sequence is produced according to the aforementioned formula.

(2) Cauchy mutation disturbance

Based on the original STOA, we set a certain probability to perform a position update using Cauchy mutation disturbance. This enhances the algorithm's ability to escape local optima. The formula for the Cauchy probability density function is as follows:

$$f(x; x_0, \lambda) = \frac{1}{\pi \lambda \left[ 1 + \left( \frac{x - x_0}{\lambda} \right)^2 \right]} = \frac{1}{\pi} \left[ \frac{\lambda}{(x - x_0)^2 + \lambda^2} \right]$$

Incorporating this into the STOA position update formula, we have:

$$X_{newbest} = X_{best} + X_{best} \times Cauchy(0,1)$$

where  $Cauchy()$  represents the Cauchy probability density function,  $X_{newbest}$  is the position after mutation, and  $X_{best}$  is the best position before mutation.

## 2.4 Software system development

A visualization prediction system was built based on the constructed predictive model. This system was developed using MATLAB R2022a and designed using the APP Designer functionality, resulting in an initial \*.mlapp file. Subsequently, the \*.mlapp file was compiled into an executable \*.exe file that can run independently without the need for the MATLAB environment. As long as the computer has MATLAB Runtime installed, the software can be run, effectively reducing the software's runtime environment requirements and improving its portability.

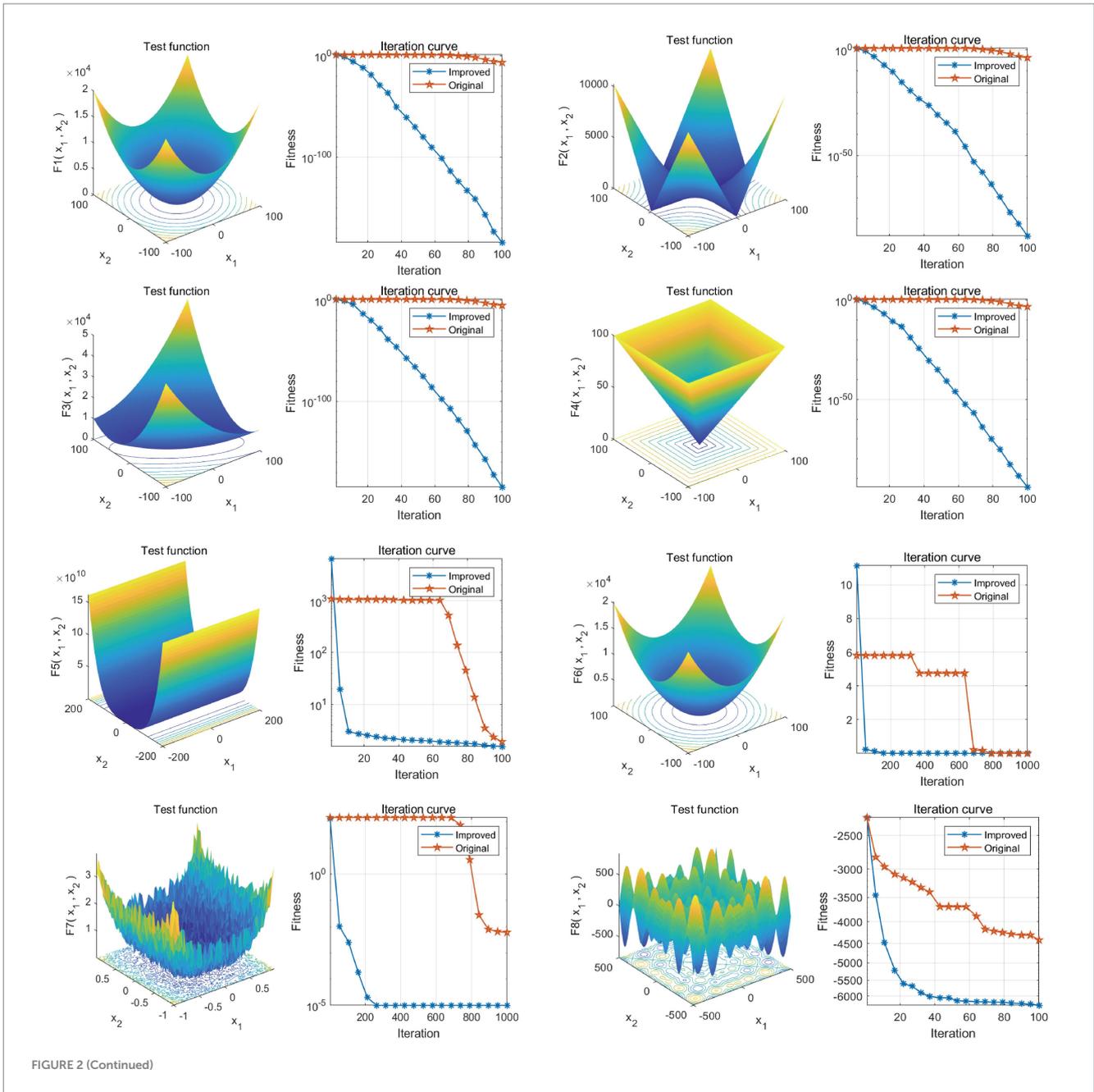
## 2.5 Statistical analysis

The predictive model construction was performed using MATLAB 2022a, and data analysis was conducted using SPSS 26.0 software. A significance level of  $p < 0.05$  was used to indicate statistically significant differences. Count data were presented as [n (%)] and compared using the chi-square test, while normally distributed continuous data were expressed as (mean ± standard deviation) and compared using the t-test.

## 3 Results

### 3.1 Performance testing of ISTOA optimization

To comprehensively evaluate the performance and efficiency of various algorithms, this study utilized 23 standard benchmark functions to assess the performance of each algorithm, the specific 23 functions are shown in [Supplementary File 1](#). The results demonstrated that ISTOA exhibited significantly faster convergence speed and



superior global optimization capability compared to other algorithms. Please refer to Figure 2a for detailed findings. Thus, the ISTOA algorithm showed clear advantages in optimization, making it suitable as the guiding algorithm for feature automatic selection in this study.

### 3.2 Lung nodule prediction model construction

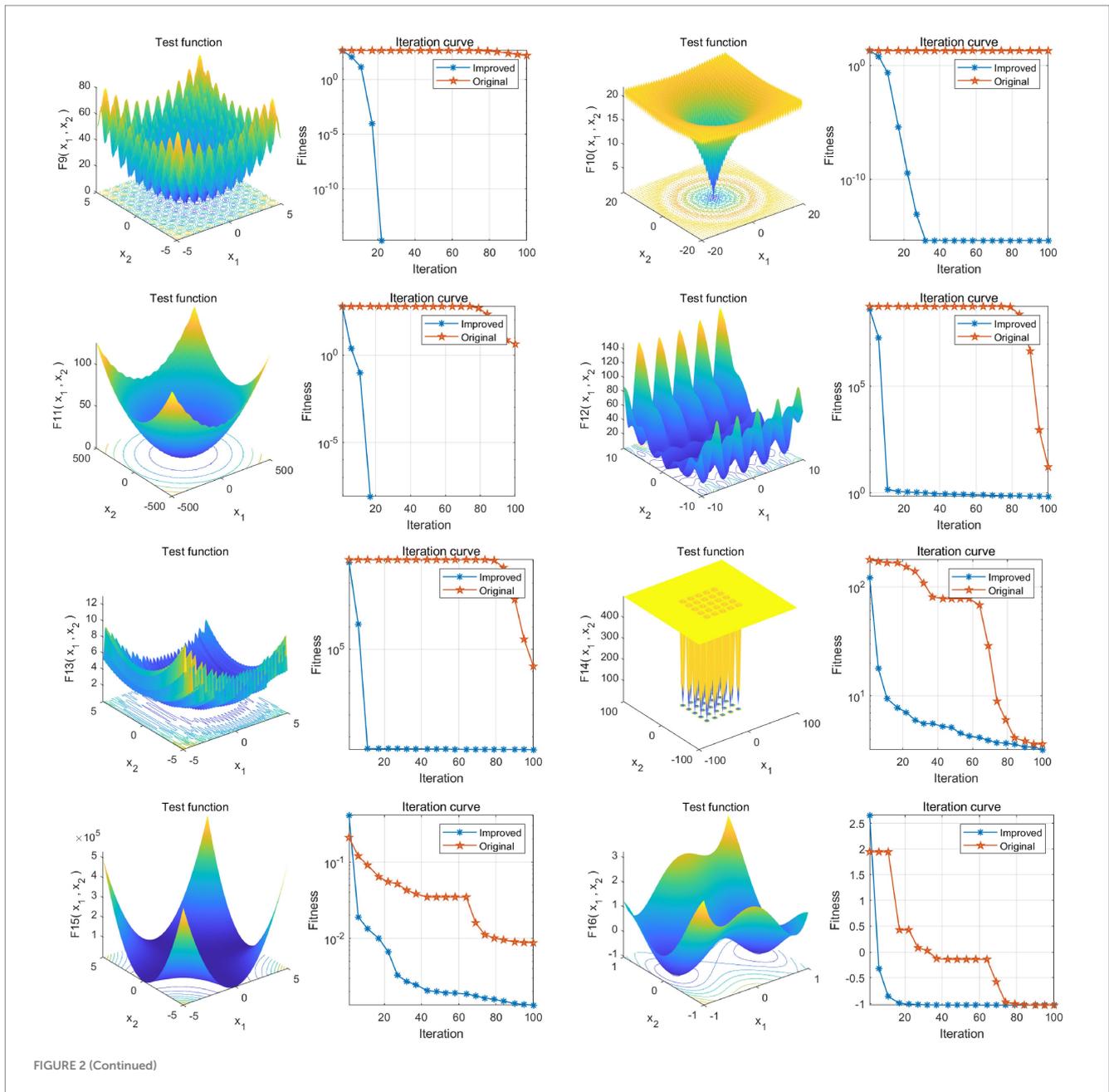
#### 3.2.1 Model construction overview

Randomly selecting 80% of the dataset as the training set (3,889 cases), we chose several base models including logistic regression (LR), decision tree (DT), k-nearest neighbors algorithm (KNN), backpropagation neural network (BP), support vector machine

(SVM), random forest (RF), and XGBoost. All these models were integrated with ISTOA for automatic feature selection, thus constructing the FSRML model. Five-fold cross-validation was performed for all models. Comparing the validation results of the five-fold cross-validation, it was evident that XGBoost exhibited significant advantages (Table 1; Figures 2b,c).

#### 3.2.2 Machine learning model performance testing

After constructing the models, the remaining 20% of the samples (972 cases) were selected as the test set to evaluate the predictive performance of each model on external data. The results showed that using XGBoost as the base model for FSRML yielded significant improvements in predictive performance (Table 2; Figures 3, 4).



### 3.2.3 Compared with other automatic machine learning methods

The FSRML we developed was compared with other AutoML models, and the results showed that the FSRML model constructed in this study had the best prediction performance on the test set (Table 3; Figures 5, 6).

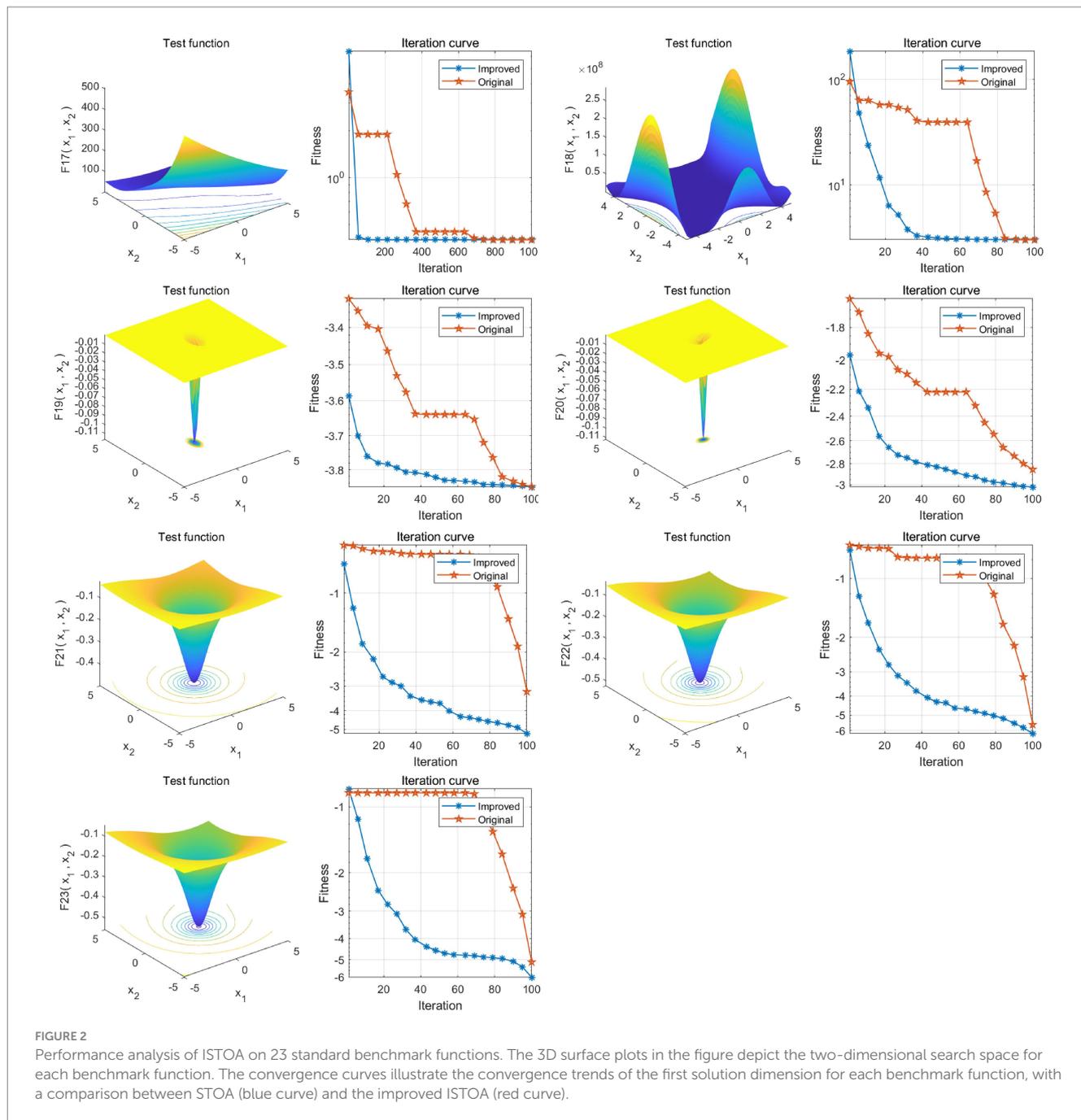
### 3.3 Feature validation through model-automated selection

The features selected automatically by the FSRML model include age, smoking or frequent passive smoking, significant psychological stress in the past year, occupational exposure (presence of air pollution in the work environment), presence of chronic lung disease, family

history of lung cancer, elevated levels of albumin, and elevated levels of carcinoembryonic antigen. The value of these selected features was assessed through univariate analysis and multivariate logistic stepwise regression analysis.

#### 3.3.1 Univariate analysis for feature selection

The results of the univariate analysis showed that in patients with positive lung nodules, the proportions of age, smoking or frequent passive smoking, significant psychological stress in the past year, occupational exposure (presence of air pollution in the work environment), presence of chronic lung disease, family history of lung cancer, and elevated levels of carcinoembryonic antigen were higher compared to patients with negative lung nodules. Additionally, the level of albumin was lower in patients with positive lung nodules. These differences were statistically significant ( $p < 0.05$ ) (Table 4).

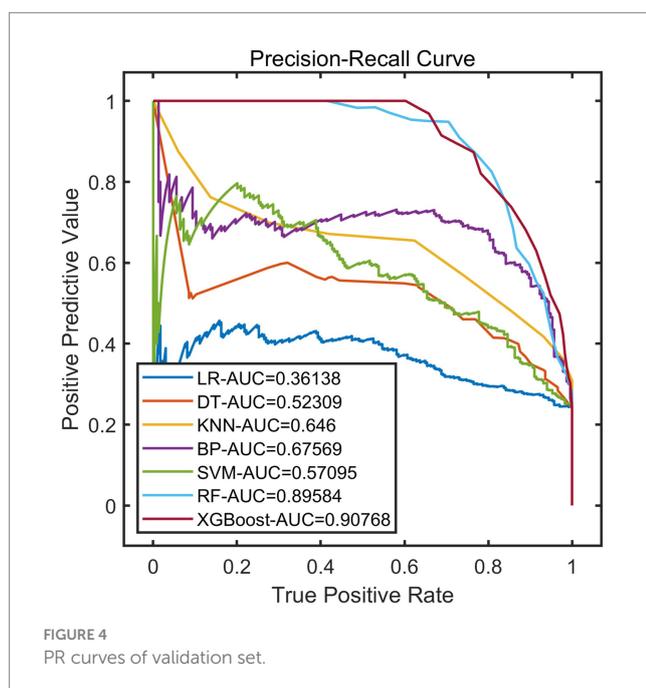
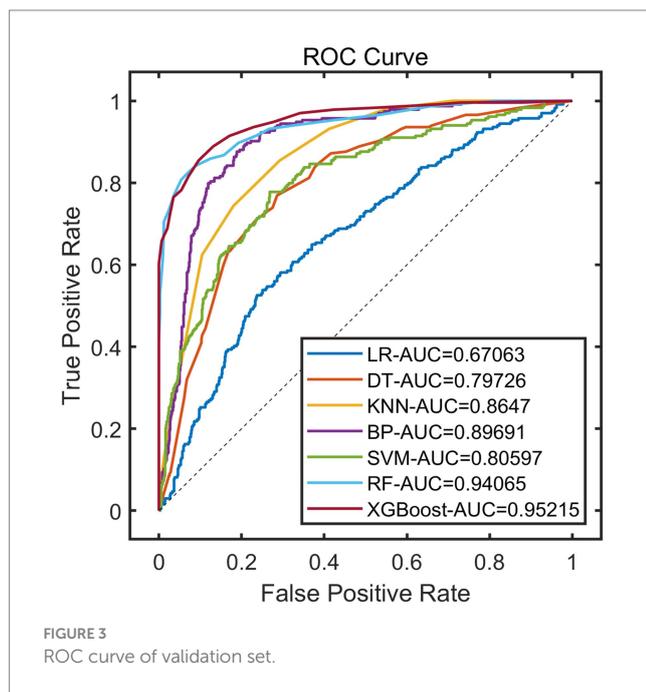


**TABLE 1** Cross-validation results of FSRML model training (validation set).

Basic model	PRE	SEN	SPE	ACC	F1	ROC-AUC	PR-AUC
LR	0.4016	0.0525	0.9753	0.7537	0.0928	0.6866	0.3681
DT	0.5602	0.4979	0.8765	0.7855	0.5272	0.7801	0.5379
KNN	0.5519	0.6660	0.8291	0.7899	0.6036	0.8480	0.6187
BP	0.6779	0.7548	0.8866	0.8550	0.7143	0.8852	0.7247
SVM	0.6131	0.4497	0.9103	0.7997	0.5188	0.8118	0.5504
RF	0.8893	0.6970	0.9726	0.9064	0.7815	0.9323	0.8726
XGBoost	0.9410	0.6831	0.9865	0.9136	0.7916	0.9496	0.9028

TABLE 2 FSRML model performance comparison results (test set).

Basic model	PRE	SEN	SPE	ACC	F1	ROC-AUC	PR-AUC
LR	0.3030	0.0427	0.9688	0.7459	0.0749	0.6706	0.3614
DT	0.5490	0.5983	0.8442	0.7850	0.5726	0.7973	0.5231
KNN	0.6547	0.6239	0.8957	0.8302	0.6389	0.8647	0.6460
BP	0.7017	0.7137	0.9038	0.8580	0.7076	0.8969	0.6757
SVM	0.6575	0.4103	0.9322	0.8066	0.5053	0.8060	0.5710
RF	0.9483	0.7051	0.9878	0.9198	0.8088	0.9407	0.8958
XGBoost	0.9148	0.6880	0.9797	0.9095	0.7854	0.9522	0.9077



### 3.3.2 Multivariate analysis for feature selection

The results of the multivariate analysis showed that advanced age, smoking or frequent passive smoking, significant psychological stress in the past year, occupational exposure (presence of air pollution in the work environment), presence of chronic lung disease, and elevated levels of carcinoembryonic antigen were identified as risk factors for predicting the occurrence of lung nodules (Table 5).

### 3.4 Development of visualization system

In clinical practice, the changes in various features related to lung nodules can be complex and difficult to visually interpret, making it challenging to determine whether a patient is at risk of developing lung nodules. Existing artificial intelligence methods also face the challenge of high implementation barriers, requiring clinicians to possess advanced coding skills and extensive literature review, which hinders widespread adoption in hospitals. To address this issue, this study innovatively developed a practical visualization system called “A Risk Prediction System for Pulmonary Nodules in Physical Examination Population.” This system offers intuitive, convenient, and practical advantages.

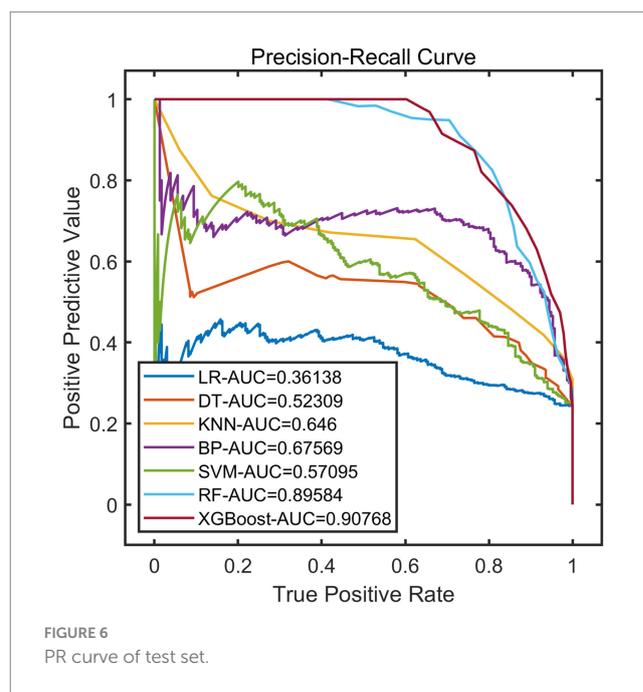
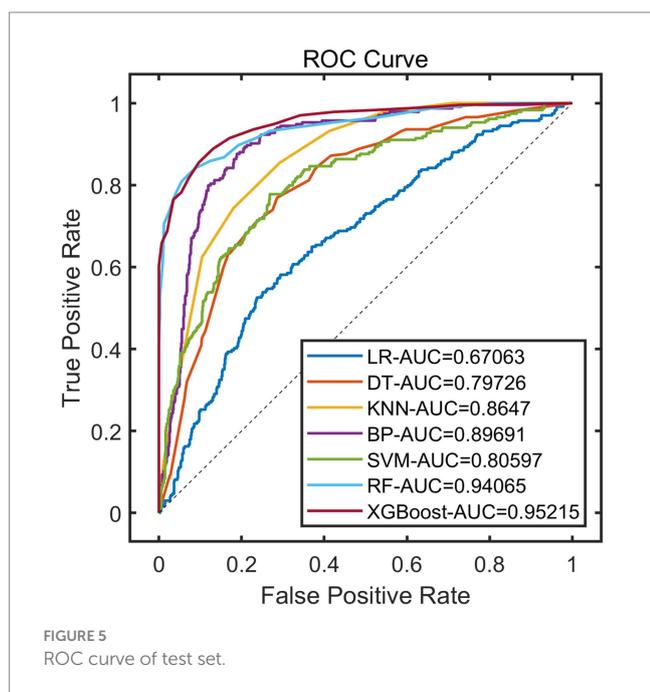
Users can input patients’ basic information in the “Basic Information Input” section and then click the “Prediction” button. The predicted results will be displayed in the “Prediction Result Output” section, providing users with easy access to the prediction outcomes (Figures 7, 8).

## 4 Discussion

In recent years, with changes in people’s lifestyles and the influence of environmental factors, the incidence of lung nodules, as one of the early signs of lung cancer, has been increasing year by year (18). Currently, the diagnostic techniques for lung nodules mainly include CT scanning, needle biopsy, or pathological examination after surgery. CT scans rely on comprehensive analysis by physicians of lesion location, size, density, shape, and other information to make a qualitative diagnosis. However, different pathological subtypes of lung nodules often exhibit similar imaging features, and the diagnosis of the same lesion may be influenced by subjective differences among different diagnosticians, making it difficult to accurately diagnose early-stage lung cancer (19, 20). Pathological

TABLE 3 Comparison of FSRML and other AutoML prediction performance.

Model	PRE	SEN	SPE	ACC	F1	ROC-AUC	PR-AUC
TPE-GP	0.7778	0.0598	0.9946	0.7695	0.1111	0.7112	0.4535
TPOT	0.9412	0.2735	0.9946	0.8210	0.4238	0.9242	0.8220
AutoSklearn	0.3529	0.0256	0.9851	0.7541	0.0478	0.6865	0.3637
AutoGluon	0.6329	0.6410	0.8821	0.8241	0.6369	0.8431	0.6755
FSRML	0.9148	0.6880	0.9797	0.9095	0.7854	0.9522	0.9077



examination is considered the gold standard for diagnosing lesions, but it does not provide a comprehensive assessment of the entire lesion. Therefore, different pathological results can also occur depending on the site of sample collection (21). Positron Emission Tomography/Computed Tomography (PET/CT), the significant role in the evaluation and management of pulmonary nodules. PET/CT is instrumental in distinguishing between benign and malignant nodules, enhancing the diagnostic accuracy beyond what is achievable with CT alone. This imaging modality integrates metabolic and anatomic information, providing a more comprehensive assessment of nodule activity. Studies have shown that PET/CT can significantly improve the sensitivity and specificity of lung cancer detection, especially in nodules that are indeterminate in size and appearance (22). However, the high cost of PET/CT makes it difficult to promote it in clinical practice.

Machine learning is an interdisciplinary field that combines statistics, various domains of knowledge, and computer technology to process large volumes of data. It is a subfield of artificial intelligence. By utilizing machine learning algorithms, researchers can extract the necessary feature variables from massive datasets, thereby enhancing learning efficiency (23, 24). Machine learning has been widely applied in the medical field. In this study, a machine learning model based on XGBoost was developed for feature

recognition. This model automates the preliminary work of machine learning, including data preparation, encoding, feature selection/extraction, and engineering environment. During the model generation process, it involves algorithm selection, optimization, iteration, and validation (25, 26). Additionally, this study utilized the ISTOA for optimizing the performance of machine learning. This algorithm builds upon the decision tree algorithm and continuously improves precision through accumulation (27). An essential aspect of implementing data-driven models in medicine is ensuring the feasibility and integration of these processes within healthcare service providers. The successful deployment of our feature self-recognition machine learning model for pulmonary nodule risk screening hinges not only on its predictive accuracy but also on its practical application in clinical settings. According to recent studies, it is crucial to consider factors such as interoperability with existing healthcare systems, ease of use for clinical staff, and the ability to handle large-scale data efficiently. Our model has been designed with these considerations in mind, featuring an intuitive visualization system that can seamlessly integrate with electronic health records (EHR) and other hospital information systems (HIS). Additionally, the model's reliance on routinely collected clinical data ensures that its implementation does not require significant changes to current workflows, thereby facilitating its adoption in real-world healthcare environments. Future work will

TABLE 4 Results of univariate analysis of model selection feature.

Feature	Negative (n = 3,693)	Positive (n = 1,168)	t/ $\chi^2$ value	p-value
Age(year)	33.87 ± 8.43	43.32 ± 9.16	32.691	<0.001
Smoking or frequent passive smoking			320.600	<0.001
Yes	1,469(39.78)	815(69.78)		
No	2,224(60.22)	353(30.22)		
Significant psychological stress in the past year			29.163	<0.001
Yes	775(20.99)	327(28.66)		
No	2,918(79.01)	814(71.34)		
Occupational exposure (presence of air pollution in the work environment)			209.891	<0.001
Yes	443(12.0)	350(29.97)		
No	3,250(88.0)	818(70.03)		
Presence of chronic lung disease			41.936	<0.001
Yes	187(5.06)	121(10.36)		
No	3,506(94.94)	1,047(89.64)		
Family history of lung cancer			136.618	<0.001
Yes	517(14.0)	338(28.94)		
No	3,176(86.0)	830(71.06)		
Albumin (g/L)	49.14 ± 11.23	44.92 ± 9.33	11.635	<0.001
Elevated levels of carcinoembryonic antigen			40.807	<0.001
Yes	61(1.65)	58(4.97)		
No	3,632(98.35)	1,110(95.03)		

TABLE 5 Results of multivariate logistic regression stepwise regression analysis.

Feature	$\beta$	SE	Wald	P	EXP(B)	95% CI for EXP(B)	
						Lower	Upper
Age (year)	1.536	0.636	5.833	0.019	4.646	3.399	5.893
Smoking or frequent passive smoking	1.231	0.311	15.667	<0.001	3.425	2.815	4.034
Significant psychological stress in the past year	0.515	0.134	14.771	<0.001	1.674	1.411	1.936
Presence of chronic lung disease	0.742	0.237	9.802	0.003	2.100	1.636	2.565
Elevated levels of carcinoembryonic antigen	1.011	0.352	8.249	0.006	2.748	2.058	3.438
Occupational exposure (presence of air pollution in the work environment)	1.067	0.491	4.722	0.034	2.907	1.944	3.869

focus on pilot testing the system in various healthcare settings to further validate its feasibility and gather feedback for continuous improvement.

In this study, both univariate analysis and multivariate logistic regression models were used to identify six influencing factors: advanced age, smoking or frequent passive smoking, significant psychological stress in the past year, occupational exposure (presence of air pollution in the work environment), presence of chronic lung disease, and elevated levels of carcinoembryonic antigen. On the other hand, the feature recognition machine learning model identified eight features, including age, smoking or frequent passive smoking, significant psychological stress in the past year, occupational exposure (presence of air pollution in the work environment), presence of chronic lung disease, family history of lung cancer, decreased levels of

albumin, and elevated levels of carcinoembryonic antigen. These features can be used for early diagnosis and prediction of the risk of developing lung nodules. This is because in the regression models, there is a high degree of linear correlation among the independent variables, which leads to inaccurate, unstable, and even unreliable estimates of the regression coefficients. This affects the predictive ability of the models and indicates that machine learning outperforms traditional multivariate analysis.

An analysis of the aforementioned risk factors reveals that both men and women have an increased incidence of lung nodules with age. This is because as the body ages, the immune system weakens, cell self-repair capabilities decline, and various carcinogenic factors stimulate the development of multiple diseases, promoting tumor growth (28, 29). Smoking intensity and duration are positively correlated with the

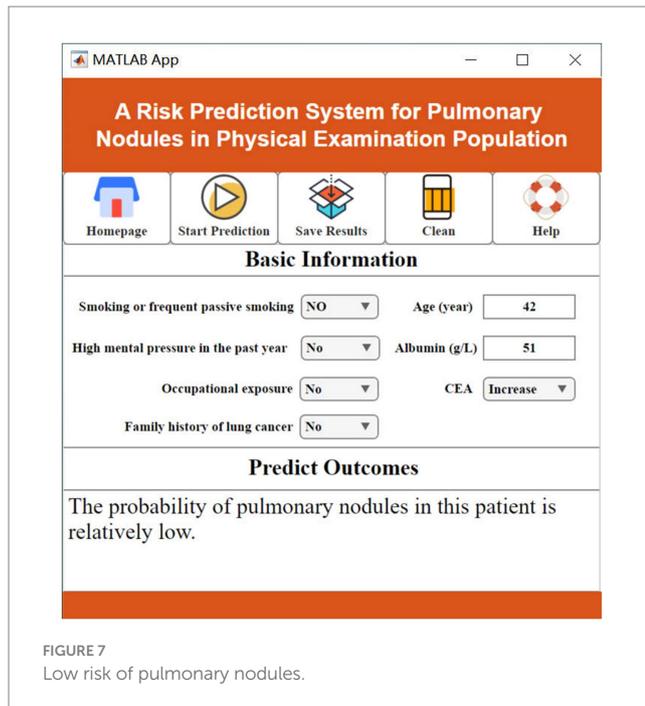


FIGURE 7  
Low risk of pulmonary nodules.

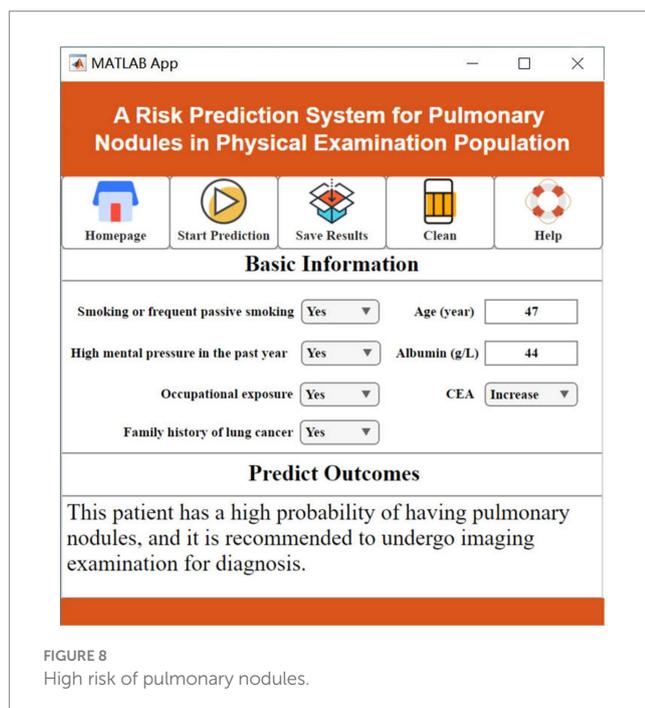


FIGURE 8  
High risk of pulmonary nodules.

incidence of lung nodules. This is due to the presence of dozens of carcinogens in tobacco, which can cause genetic mutations and promote chronic tumor growth. Passive smokers unknowingly inhale smoke, leading to lung function impairment (30, 31). Smoking also causes constriction of small blood vessels in the lungs and thickening of vessel walls, resulting in elevated levels of carcinoembryonic antigen (32). Air pollution and smoking have a synergistic effect on the occurrence and development of lung cancer, continuously increasing the incidence of lung nodules. Previous studies have shown that exposure to kitchen fumes and occupational exposure increase the risk

of developing lung nodules. This is because kitchen fumes mainly contain carcinogens such as benzopyrene, volatile nitrosamines, and heterocyclic amine compounds, which exert cytotoxic effects on lung tissue and damage the respiratory system. Occupational exposure to substances like aluminum, arsenic, asbestos, coke, and coal gas has carcinogenic effects on the lungs (33, 34). Most lung nodules are caused by lung inflammation, and underlying lung diseases such as pneumonia, emphysema, chronic bronchitis, chronic obstructive pulmonary disease (COPD), and asthma are all inflammatory conditions that can recur and increase the incidence of lung nodules (35, 36). A positive family history of lung cancer and a history of lung disease are positively associated with the development of lung nodules, increasing the risk of their occurrence (37). Reasons for low albumin levels include inadequate intake, excessive consumption, excessive elimination, and insufficient synthesis. In patients with lung nodules, low albumin levels can be caused by malnutrition, impaired liver function, tumor metastasis, digestive tract tumors, liver tumors, and other factors (38, 39).

Compared to traditional statistical models such as logistic regression, the feature recognition machine learning model improves model accuracy. Additionally, the use of the ISTOA significantly reduces the barrier to entry for artificial intelligence technology. Healthcare professionals can utilize this tool to screen individuals undergoing medical examinations for their risk of developing lung nodules and implement targeted intervention measures. For example, they can promote a balanced diet, encourage physical exercise, foster healthy lifestyle habits, and establish regulations to restrict smoking (40, 41).

## 5 Conclusion

The application of feature recognition machine learning models can help clinicians identify characteristics of lung nodule patients, thereby enabling early prediction of disease occurrence, assisting in the development of treatment plans, and improving prognosis. However, this study also has certain limitations. Firstly, it is a retrospective and single-center study, which may introduce selection bias and affect the accuracy of the research findings. To further validate the results, more multicenter sample data is needed. Additionally, CT imaging plays a crucial role in the diagnosis of lung nodules in clinical practice. However, this study did not include CT imaging radiomics features, indicating the need for further analysis in future studies.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by the Ethics Committee of the Western Theater General Hospital of the People's Liberation Army. The studies were conducted in accordance with the

local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because this study was conducted retrospectively, informed consent was waived.

## Author contributions

KH: Conceptualization, Methodology, Project administration, Resources, Writing – review & editing. FT: Data curation, Investigation, Resources, Writing – original draft. YL: Data curation, Investigation, Writing – review & editing. LW: Investigation, Writing – review & editing. FF: Formal analysis, Methodology, Software, Visualization, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## References

- Godoy MCB, Odisio EGLC, Truong MT, de Groot PM, Shroff GS, Erasmus JJ. Pulmonary nodule Management in Lung Cancer Screening: a pictorial review of lung-RADS version 1.0. *Radiol Clin North Am.* (2018) 56:353–63. doi: 10.1016/j.rcl.2018.01.003
- De Margerie-Mellon C, Chassagnon G. Artificial intelligence: a critical review of applications for lung nodule and lung cancer. *Diagn Interv Imaging.* (2023) 104:11–7. doi: 10.1016/j.diii.2022.11.007
- Wu Z, Wang F, Cao W, Qin C, Dong X, Yang Z, et al. Lung cancer risk prediction models based on pulmonary nodules: a systematic review. *Thorac Cancer.* (2022) 13:664–77. doi: 10.1111/1759-7714.14333
- Zheng F, Lavin J, Sprafka JM. Patient out-of-pocket costs for suspicious pulmonary nodule biopsy in lung cancer patients. *J Med Econ.* (2021) 24:1173–7. doi: 10.1080/13696998.2021.1988282
- Zhang Y, Jiang B, Zhang L, Greuter MJW, de Bock GH, Zhang H, et al. Lung nodule detectability of artificial intelligence-assisted CT image Reading in lung Cancer screening. *Curr Med Imaging.* (2022) 18:327–34. doi: 10.2174/1573405617666210806125953
- Ghossein J, Gingras S, Zeng W. Differentiating primary from secondary lung cancer with FDG PET/CT and extra-pulmonary tumor grade. *J Med Imaging Radiat Sci.* (2023) 54:451–6. doi: 10.1016/j.jmir.2023.05.045
- Li W, Yu S, Yang R, Tian Y, Zhu T, Liu H, et al. Machine learning model of ResNet50-ensemble voting for malignant-benign small pulmonary nodule classification on computed tomography images. *Cancers (Basel).* (2023) 15:5417. doi: 10.3390/cancers15225417
- Liu M, Zhou Z, Liu F, Wang M, Wang Y, Gao M, et al. CT and CEA-based machine learning model for predicting malignant pulmonary nodules. *Cancer Sci.* (2022) 113:4363–73. doi: 10.1111/cas.15561
- Pei Q, Luo Y, Chen Y, Li J, Xie D, Ye T. Artificial intelligence in clinical applications for lung cancer: diagnosis, treatment and prognosis. *Clin Chem Lab Med.* (2022) 60:1974–83. doi: 10.1515/cclm-2022-0291
- Chassagnon G, de Margerie-Mellon C, Vakalopoulou M, Marini R, Hoang-Thi TN, Revel MP, et al. Artificial intelligence in lung cancer: current applications and perspectives. *Jpn J Radiol.* (2023) 41:235–44. doi: 10.1007/s11604-022-01359-x
- Goncalves S, Fong PC, Blokhina M. Artificial intelligence for early diagnosis of lung cancer through incidental nodule detection in low- and middle-income countries-acceleration during the COVID-19 pandemic but here to stay. *Am J Cancer Res.* (2022) 12:1–16.
- Viswanathan VS, Toro P, Corredor G, Mukhopadhyay S, Madabhushi A. The state of the art for artificial intelligence in lung digital pathology. *J Pathol.* (2022) 257:413–29. doi: 10.1002/path.5966
- Zhang K, Chen K. Artificial intelligence: opportunities in lung cancer. *Curr Opin Oncol.* (2022) 34:44–53. doi: 10.1097/CCO.0000000000000796

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2024.1424750/full#supplementary-material>

- Xia Q, Ding Y, Zhang R, Zhang H, Li S, Li X. Optimal performance and application for seagull optimization algorithm using a hybrid strategy. *Entropy.* (2022) 24:973. doi: 10.3390/e24070973
- Yang C, Pan P, Ding Q. Image encryption scheme based on mixed chaotic Bernoulli measurement matrix block compressive sensing. *Entropy.* (2022) 24:273. doi: 10.3390/e24020273
- Araújo MO, Marinho LS, Felinto D. Observation of nonclassical correlations in biphotons generated from an ensemble of pure two-level atoms. *Phys Rev Lett.* (2022) 128:83601. doi: 10.1103/PhysRevLett.128.083601
- Zhi Z, Bian Z, Chen Y, Zhang X, Wu Y, Wu H. Horizontal and vertical comparison of microbial community structures in a low permeability reservoir at the local scale. *Microorganisms.* (2023) 11:2862. doi: 10.3390/microorganisms11122862
- Senent-Valero M, Librero J, Pastor-Valero M. Solitary pulmonary nodule malignancy predictive models applicable to routine clinical practice: a systematic review. *Syst Rev.* (2021) 10:308. doi: 10.1186/s13643-021-01856-6
- Silva M, Milanese G, Sestini S, Sabia F, Jacobs C, van Ginneken B, et al. Lung cancer screening by nodule volume in lung-RADS v1.1: negative baseline CT yields potential for increased screening interval. *Eur Radiol.* (2021) 31:1956–68. doi: 10.1007/s00330-020-07275-w
- Ha T, Kim W, Cha J, Lee YH, Seo HS, Park SY, et al. Differentiating pulmonary metastasis from benign lung nodules in thyroid cancer patients using dual-energy CT parameters. *Eur Radiol.* (2022) 32:1902–11. doi: 10.1007/s00330-021-08278-x
- Wang Y, Huang Q, Li J. Analysis of clinical and pathological features of malignant pulmonary nodules. *Altern Ther Health Med.* (2023) 29:188–93.
- Evangelista L, Cuocolo A, Pace L, Mansi L, del Vecchio S, Miletto P, et al. Performance of FDG-PET/CT in solitary pulmonary nodule based on pre-test likelihood of malignancy: results from the ITALIAN retrospective multicenter trial. *Eur J Nucl Med Mol Imaging.* (2018) 45:1898–907. doi: 10.1007/s00259-018-4016-1
- Zhang Y, Feng W, Wu Z, Li W, Tao L, Liu X, et al. Deep-learning model of ResNet combined with CBAM for malignant-benign pulmonary nodules classification on computed tomography images. *Medicina.* (2023) 59:1088. doi: 10.3390/medicina59061088
- Chen Y, Hou X, Yang Y, Ge Q, Zhou Y, Nie S. A novel deep learning model based on multi-scale and multi-view for detection of pulmonary nodules. *J Digit Imaging.* (2023) 36:688–99. doi: 10.1007/s10278-022-00749-x
- Huang W, Zhang H, Ge Y, Duan S, Ma Y, Wang X, et al. Radiomics-based machine learning methods for volume doubling time prediction of pulmonary ground-glass nodules with baseline chest computed tomography. *J Thorac Imaging.* (2023) 38:304–14. doi: 10.1097/RTI.0000000000000725
- Qi J, Hong B, Tao R, Sun R, Zhang H, Zhang X, et al. Prediction model for malignant pulmonary nodules based on cfMedIP-seq and machine learning. *Cancer Sci.* (2021) 112:3918–23. doi: 10.1111/cas.15052

27. Lin RY, Zheng YN, Lv FJ, Fu BJ, Li WJ, Liang ZR, et al. A combined non-enhanced CT radiomics and clinical variable machine learning model for differentiating benign and malignant sub-centimeter pulmonary solid nodules. *Med Phys.* (2023) 50:2835–43. doi: 10.1002/mp.16316
28. Xu L, Su Z, Xie B. Diagnostic value of conventional tumor markers in young patients with pulmonary nodules. *J Clin Lab Anal.* (2021) 35:23912. doi: 10.1002/jcla.23912
29. Huang C, Sun Y, Wu Q, Ma C, Jiao P, Wang Y, et al. Simultaneous bilateral pulmonary resection via single-utility port VATS for multiple pulmonary nodules: a single-center experience of 16 cases. *Thorac Cancer.* (2021) 12:525–33. doi: 10.1111/1759-7714.13791
30. Gendarme S, Chouaid C. Monitoring subsolid pulmonary nodules in high-risk patients is even more cost-effective when combined with a stop-smoking program. *J Thorac Oncol.* (2020) 15:1268–70. doi: 10.1016/j.jtho.2020.04.023
31. Trejo Gallego C, Bueno J, Cruces E, Stelow EB, Mancheño N, Flors L. Pulmonary histiocytosis: beyond Langerhans cell histiocytosis related to smoking. *Radiologia.* (2019) 61:215–24. doi: 10.1016/j.rxeng.2019.03.004
32. Huang CS, Chen CY, Huang LK, Wang WS, Yang SH. Prognostic value of postoperative serum carcinoembryonic antigen levels in colorectal cancer patients who smoke. *PLoS One.* (2020) 15:233687. doi: 10.1371/journal.pone.0233687
33. Hirano T, Numakura T, Moriyama H, Saito R, Shishikura Y, Shiihara J, et al. The first case of multiple pulmonary granulomas with amyloid deposition in a dental technician; a rare manifestation as an occupational lung disease. *BMC Pulm Med.* (2018) 18:77. doi: 10.1186/s12890-018-0654-0
34. Hung SC, Wang YT, Tseng MH. An interpretable three-dimensional artificial intelligence model for computer-aided diagnosis of lung nodules in computed tomography images. *Cancers.* (2023) 15:4655. doi: 10.3390/cancers15184655
35. Namireddy MK, Consul N, Sher AC. FDG-avid pulmonary nodules and tracheobronchial mural inflammation in IgG4-related disease. *Clin Nucl Med.* (2021) 46:e125–6. doi: 10.1097/RLU.0000000000003358
36. Liao J, Guan H, Yu M, Zhou P, Han Y, Peng X, et al. Pulmonary granulomatous inflammation after ceritinib treatment in advanced ALK-rearranged pulmonary adenocarcinoma. *Investig New Drugs.* (2022) 40:1141–5. doi: 10.1007/s10637-022-01270-2
37. Uthoff JM, Mott SL, Larson J, Neslund-Dudas CM, Schwartz AG, Sieren JC. Computed tomography features of lung structure have utility for differentiating malignant and benign pulmonary nodules. *Chronic Obstr Pulm Dis.* (2022) 9:154–64. doi: 10.15326/jcopdf.2021.0271
38. Wei Q, Fang W, Chen X, Yuan Z, du Y, Chang Y, et al. Establishment and validation of a mathematical diagnosis model to distinguish benign pulmonary nodules from early non-small cell lung cancer in Chinese people. *Transl Lung Cancer Res.* (2020) 9:1843–52. doi: 10.21037/tlcr-20-460
39. Dailey WA, Frey GT, McKinney JM, Paz-Fumagalli R, Sella DM, Toskich BB, et al. Percutaneous computed tomography-guided radiotracer-assisted localization of difficult pulmonary nodules in Uniportal video-assisted thoracic surgery. *J Laparoendosc Adv Surg Tech A.* (2018) 28:1451–7. doi: 10.1089/lap.2018.0248
40. Kao MW. Intracorporeal direct measurement for localizing peripheral pulmonary nodules during thoracoscopy. *J Thorac Dis.* (2019) 11:4119–26. doi: 10.21037/jtd.2019.10.06
41. Kim H, Goo JM, Park CM. A simple prediction model using size measures for discrimination of invasive adenocarcinomas among incidental pulmonary subsolid nodules considered for resection. *Eur Radiol.* (2019) 29:1674–83. doi: 10.1007/s00330-018-5739-x