Indian Journal of Experimental Biology
Vol. 61, July 2023, pp. 485-492
DOI: 10.56042/ijeb.v61i07.2826

Machine learning to predict etiology for infectious diseases of classic fever of unknown origin in adults

Yani Zhou[1], Cha Chen[1], Bing Ruan[2] & Weihong Wang[1,3]*

[1]Department of Infectious Diseases, Huzhou Central Hospital, Affiliated Central Hospital Huzhou University, Affiliated Huzhou Hospital, Zhejiang University School of Medicine, Huzhou-313000, Zhejiang Province, China

[2]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, National Medical Center for Infectious Diseases, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou-310003, Zhejiang Province, China

[3]Huzhou Key Laboratory of Precision Medicine Research and Translation for Infectious Diseases, Huzhou, Zhejiang Province, China

Received 19 May 2023; *revised 14 June 2023*

The etiologies of infectious diseases (IDs) of classic fever of unknown origin (FUO) are multitudinous. Different etiologies affect medication decisions. Here, we have made an attempt to predict the types of etiology on the basis of a machine learning (ML) model for IDs of classic FUO for adults. Ten years clinical data of 408 classic FUO were retrospectively collected from August 2012 to August 2022 in Huzhou Central Hospital. A total of 256 adult patients with ID of classic FUO were divided into four subgroups for clinical characteristic analysis. Random forest (RF), light gradients boosting (Light GBM), and extreme gradient boosting (XGBoost) were used to construct prediction models of 10-fold cross-validation. The micro average and weighted average of F1 score were calculated to evaluate the performance of the models. SHapley Additive exPlanations (SHAP) was used to explain the relationship between features and the predicted results. Clinical characteristic analysis showed that 25 indices were statistically different (P <0.05). RF, LightGBM and XGBoost models, were constructed on the basis of these indices. Among them, the XGBoost model showed the best performance (micro-F1=0.7129，weighted-F1=0.6618). The areas under the ROC curve of the four subgroups were 0.7477, 0.7162, 0.9200 and 0.7500, respectively. C-reactive protein (CRP), N%, C3, and C4 with high SHAP values were positively correlated with the bacterial ID model output, which was used to distinguished other causes. Bacterial infections were the main cause of IDs. The XGBoost model could be regarded as an auxiliary tool to predict the etiological types of IDs of classic FUO, improve the etiological diagnostic rate, and provide evidence for clinical drug application.

Keywords: C-reactive protein (CRP), Extreme gradient boosting (XGBoost), Light gradients boosting (Light GBM), Random forest (RF), SHAP

Classic FUO involves a wide range of etiologies. Diagnosing diseases is a great challenge for clinicians[1]. The classic FUO distributed in the community, outpatients, or inpatients is the most common type of FUO. Classic FUO is a group of diseases with the following characteristics: (i) fever >38.3°C for more than 3 weeks; and (ii) diagnosis that cannot be confirmed after systematic and comprehensive examination[2,3].

The etiology of classic FUO can be summarized into four types: infectious diseases (IDs), neoplastic diseases, non-infectious inflammatory diseases (NIIDs), and other diseases[4]. The etiology of IDs can also be divided into bacterial infection, viral infection, fungal infection, and other pathogens. In addition to conventional pathogen detection methods, recent studies have shown that 18F-fluorodeoxyglucose positron emission tomography and CT[5], next-generation sequencing (NGS)[6,7], and molecular diagnostic assay are helpful to improve the disease diagnostic rate[8]. However, only a small number of patients could be diagnosed. Clinicians often give empirical anti-infective therapy through preliminary clinical diagnosis when the pathogen or etiology is not clear.Empirical anti-infective therapy not only causes a decrease in the positive rate of pathogen examination to delay diagnosis but also leads to huge economic waste, double infection, and drug-resistant bacteria, and its therapeutic effect has not been confirmed[9]. Therefore, how to improve the etiological diagnosis rate for IDs of classic FUO remains a difficult problem.

*Correspondence:
E-Mail: hzwwh0606@163.com

In recent years, machine learning (ML) has been widely used to predict the occurrence, development, prognosis and survival rate of diseases[10,11]. Yan *et al*.[12] has used ML to predict the etiology of classic FUO, but the ID types of classic FUO have not been further analyzed. Hence, in the present study, we explored for a ML model with high accuracy as an auxiliary tool for etiological diagnosis of ID by retrospective analysis of IDs of classic FUO to promote rational drug application.

Methodolgy

Subjects

The clinical data of 408 adults with classic FUO from August 2012 to August 2022 in Huzhou Central Hospital were collected, and 256 of them were definitively diagnosed as infectious diseases. Exclusion criteria: (i) Age <18; (ii) Pregnancy; (iii) Nosocomial infection; (iv) History of immunodeficiency, such as HIV, long-term immunosuppressant use, and allogeneic or autologous organ transplant status; (v) non-infectious inflammatory diseases (NIIDs); and (vi) Other etiologies of classic FUO. A total of 256 patients were divided into four subgroups viz. bacterial (n=195), viral (n=36), fungal (n=10); and other pathogens (n=15).

Data collection

The clinical data included general information (gender and age) and clinical data (symptoms, physical examination, medical history, and first laboratory examination on admission and discharge diagnosis).

Model development

The dataset was split into four groups to develop the models. Predictive models were built with RF, Light GBM and XGBoost, which play a key role in formulating the base line for the establishment of the model. RF, Light GBM, XGBoost are the class tree models using ensemble learning idea. The Light GBM and XGBoost are both asymmetric trees and developed methods recently, but Light GBM grows leaf-wise while XGBoost grows level-wise[13,14].

All model construction in this study was completed in Python 3.10. Install and download Numpy, Matplotlib-3.6.1, Scipy-1.9.2, Scikit_learn-1.1.2 and pandas for multi-class data analysis, cleaning, preparation, and computation. Random Forest Classifier was imported from sklearn, and XGB Classifier and LGBM Classifier were loaded from XGBoost library and LightGBM library, respectively. The accuracy of models was checked by 10-fold cross validation.

Model evaluation

Accuracy, Precision, Recall and F1-score were included to calculate the micro average and weighted average values, and then to evaluate the performance of the multi-classification models.

Statistical analysis

Statistical software SPSS25.0 (IBM Corp., Armonk, NY, U.S.A.) was used for data analysis. $\chi 2$ test and Fisher's exact test were used to compare categorical variables. Measurement data with normal distribution were expressed as mean ± standard deviation (x± s), and data with skewed distribution were expressed as median and quartile [M (P25, P75)]; The Kruskal Wallis test was used for multivariate nonparametric rank-sum test. P <0.05 was statistically significant for all difference indices.

Results

Etiological analysis of IDs of classic FUO in adults

A total of 256 patients with a definite diagnosis of IDs of classic FUO were used in this study. The etiological analysis of IDs of classic FUO included bacterial, viral, fungal and other pathogen infections. Among them, bacterial infection (76.17%) was the main cause. Moreover, 14.06%were infected with viruses, 3.91% were infected with fungi, and 5.86% were infected with other pathogens. The disease classification is shown in Table 1.

Clinical characteristic analysis of IDs of classic FUO in adults

The results of clinical characteristic analysis showed that age (P <0.001) and rash ($\chi 2$=9.829, P =0.012) were statistically significant (Table 2). Given that age is a continuous variable, its analysis results are shown in Table 3. Significant differences in age could be observed among the four subgroups, and the average age of the viral infection group was the smallest (41.64 ±14.20). In addition, 23 differential indices of laboratory examination (P <0.05) are shown in Table 3. A total of 25 differential indices were found by analyzing the clinical characteristic.

Establishment and evaluation of prediction models

RF, LightGBM, and XGBoost models were constructed on the basis of the25 differential indices screened above, and the model performance was evaluated in terms of accuracy, precision, recall, and F1-score. The micro average and weighted average values of the above four indices were calculated. The results of the RF, LightGBM and XGBoost models were as follows: the accuracy values were 0.7175,

Table 1 — Etiology analysis of infectious diseases of classic FUO in adults

Etiology		N (%)	Etiology		N (%)
Bacterial IDs		195 (76.17)	Central infection	Purulent meningitis	2 (0.78)
Blood infection	MRS	17 (6.64)	Cellulitis		1 (0.51)
Urinary infection	Urinary tract infections	20 (7.81)	Infective endocarditis		24 (9.38)
	Pyelonephritis	6 (2.34)	Brucellosis		6 (2.34)
Abscess	Liver abscess	8 (3.12)	Pseudomembranous colitis		1 (0.51)
	Tuberculous abscess	2 (0.78)	Viral IDs	36 (14.06)	Viral IDs
	Pelvic abscess	1 (0.51)	EB virus		5 (1.95)
	Iliopsoas abscess	3 (1.17)	Cytomegalovirus		4 (1.56)
	Brain abscess	2 (0.78)	HIV		
	Lung abscess	1 (0.51)	Hepatitis B virus		1 (0.51)
Spinal infection	Spinal osteomyelitis	2 (0.78)	Viral meningitis		6 (2.34)
Reproductive system	Oviduct ovarian abscess	1 (0.51)	Hantavirus		2 (0.78)
Respiratory infection	Pyothorax	1 (0.51)	Measles virus		1 (0.51)
	Pulmonary infection	63 (24.61)	varicella zoster virus		1 (0.51)
	NTM	4 (1.56)	Other virus		6 (2.34)
	Actinomycosis	3 (1.17)	Fungal IDs		10 (3.91)
Mycobacterium tuberculosis infection			*Aspergillus*	Aspergillosis	4 (1.56)
Intrapulmonary	Tuberculosis	7 (2.73)	*Cryptococcus*	Cryptococcosis	4 (1.56)
	Tuberculous pleurisy	4 (1.56)	*Pneumocystis jiroveci*		1 (0.51)
Extrapulmonary	Tubercular meningitis	6 (2.34)	*Candida*	Pneumono moniliasis	1 (0.51)
	Tuberculous pericarditis	1 (0.51)	Other pathogens		15 (5.86)
	Renal tuberculosis	1 (0.51)	Chlamydia psittaci	Psittacosis	8 (3.13)
	Tuberculous lymphadenitis	3 (1.17)	Orientia tsutsugamushi	Tsutsugamushi	4 (1.56)
Peritoneal infection		1 (0.51)	Mycoplasma	Scrub typhus	2 (0.78)
Biliary infection		2 (0.78)	HansettBartonite	Cat-scratch disease	1 (0.51)
Implant infection		1 (0.51)			

Table 2 — Clinical information analysis of 256 adults with infectious diseases of classic FUO

	Bacterial	Viral	Fungal	Other pathogens	$\chi2$	*P*
Gender(n, %)					4.220	0.237
FM	90 (46.2%)	12 (33.3%)	2 (20.0%)	6 (40.0%)		
M	105 (53.8%)	24 (66.7%)	8 (80.0%)	9 (60.0%)		
Hypertension (n, %)					3.234	0.336
No	41 (21.0%)	4 (11.1%)	3 (30.0%)	4 (26.7%)		
Yes	154 (79.0%)	32 (88.9%)	7 (70.0%)	11 (73.3%)		
Diabetes (n, %)					2.084	0.527
No	20 (10.3%)	1 (2.8%)	1 (10.0%)	1 (6.7%)		
Yes	175 (89.7%)	35 (97.2%)	9 (90.0%)	14 (93.3%)		
Oral glucocorticoids with long history (n, %)					3.818	0.420
No	1 (0.5%)	1 (2.8%)	0 (0.0%)	0 (0.0%)		
Yes	194 (99.5%)	35 (97.2%)	10 (100.0%)	15 (100.0%)		
Headache (n, %)					1.303	0.722
No	19 (9.7%)	5 (13.9%)	0 (0.0%)	1 (6.7%)		
Yes	176 (90.3%)	31 (86.1%)	10 (100.0%)	14 (93.3%)		
Throat pain (n, %)					2.341	0.560
No	2 (1.0%)	1 (2.8%)	0 (0.0%)	0 (0.0%)		
Yes	193 (99.0%)	35 (97.2%)	10 (100.0%)	15 (100.0%)		
Cough (n, %)					1.598	0.679
No	30 (15.4%)	8 (22.2%)	1 (10.0%)	3 (20.0%)		
Yes	165 (84.6%)	28 (77.8%)	9 (90.0%)	12 (80.0%)		
Expectoration (n, %)					1.991	0.499
No	15 (7.7%)	5 (13.9%)	0 (0.0%)	1 (6.7%)		
Yes	180 (92.3%)	31 (86.1%)	10 (100.0%)	14 (93.3%)		
Abdominal pain (n, %)					0.881	1.000
No	4 (2.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)		
Yes	191 (97.9%)	36 (100.0%)	10 (100.0%)	15 (100.0%)		
Joint pain (n, %)					0.543	1.000
No	5 (2.6%)	1 (2.8%)	0 (0.0%)	0 (0.0%)		
Yes	190 (97.4%)	35 (97.2%)	10 (100.0%)	15 (100.0%)		
Myodynia (n, %)					4.026	0.191
No	5 (2.6%)	3 (8.3%)	0 (0.0%)	1 (6.7%)		
Yes	190 (97.4%)	33 (91.7%)	10 (100.0%)	14 (93.3%)		
Erythra (n, %)					9.829	0.012
No	7 (3.6%)	2 (5.6%)	0 (0.0%)	4 (26.7%)		
Yes	188 (96.4%)	34 (94.4%)	10 (100.0%)	11 (73.3%)		
Lymphadenectasis (n, %)					5.826	0.110
No	2 (1.0%)	2 (5.6%)	0 (0.0%)	1 (6.7%)		
Yes	193 (99.0%)	34 (94.4%)	10 (100.0%)	14 (93.3%)		
Hepatomegaly (n, %)					6.735	0.238
No	0 (0.0%)	1 (2.8%)	0 (0.0%)	0 (0.0%)		
Yes	195 (100.0%)	35 (97.2%)	10 (100.0%)	15 (100.0%)		
Splenomegaly (n, %)					0.899	0.778
No	8 (4.1%)	1 (2.8%)	0 (0.0%)	1 (6.7%)		
Yes	187 (95.9%)	35 (97.2%)	10 (100.0%)	14 (93.3%)		

(*Contd.*)

0.6549, and 0.6782, respectively; the micro-precision values were 0.7292, 0.7347, and 0.7021, respectively; the weighted-precision values were 0.6660, 0.6779, and 0.5680, respectively; the micro-recall values were 0.6731, 0.6923, and 0.6346, respectively; the weighted-recall values were 0.6731, 0.6923, and 0.6346, respectively; the micro-F1 scores were 0.7000, 0.7129,

and 0.6667, respectively; the weighted-F1 scores were 0.6309, 0.6618, and 0.5949, respectively. Models have different predictive abilities for different subgroups. For the RF model, the AUCs of the four subgroups were 0.7153, 0.7306, 0.8700 and 0.6950. For the XGBoost model, the AUCs were 0.7477, 0.7162, 0.9200 and 0.7500. For the LightGBM model, the AUCs were 0.7081, 0.7361, 0.9000 and 0.8100. A comprehensive comparison of the prediction ability of the three models showed that the XGBoost model was relatively better than the other two models (Table 4 and Fig. 1).

Bar charts were drawn to further show the contribution degree of each index for the etiology prediction model, and SHAP was used to explain the relationship between the indices and the output of the prediction model. As shown in Fig. 2A, erythra, N%, HGB, AG, C3, hypertension, headache, T3, CRP, and

Table 3 — Laboratory examination indices analysis of 256 adults with infectious diseases of classic FUO

	Bacterial	Viral	Fungal	Other pathogens	H (K)	*P*
Age (year)	56.41±15.74	41.64±14.20	59.20±18.37	54.73±11.95		<0.001
BASO%	0.30 (0.20,0.50)	0.45(0.30,0.90)	0.35 (0.20,0.60)	0.30 (0.20,0.50)	11.623	0.009
HGB (g/L)	113.00 (97.00,128.00)	130.50(110.00,140.00)	133.00 (116.75,138.00)	119.00 (101.00,136.00)	17.236	0.001
Lymphocyte (10^9/L)	1.10 (0.80,1.50)	1.50(1.10,3.03)	1.10 (0.75,1.33)	0.80(0.60,1.40)	17.322	0.001
LYMPH%	14.30 (9.90,22.10)	27.75(20.30,42.90)	10.50 (7.78,14.78)	18.80 (9.70,31.20)	33.036	<0.001
MCHC (g/L)	325.00 (316.00,333.00)	332.00 (324.00,339.75)	332.00 (324.25,340.00)	332.00 (320.00,343.00)	9.835	0.02
Neutrophils (109/L)	5.40 (3.30,8.10)	3.20 (2.25,4.78)	8.70 (4.90,10.55)	4.70 (2.40,6.70)	24.710	<0.001
NEUT%	75.10 (66.50,82.40)	58.45 (40.08,69.40)	79.75 (75.25,85.68)	67.50 (50.40,80.90)	30.679	<0.001
RBC (10^{12}/L)	3.86 (3.37,4.32)	4.33 (3.98,4.74)	4.33 (3.67,4.71)	3.93 (3.27,4.49)	18.194	<0.001
RDW (%)	13.20 (12.50,14.20)	12.23 (12.13,13.88)	12.75 (12.28,13.95)	12.50 (12.00,12.90)	11.102	0.011
CRP (mg/L)	56.20 (20.50,91.30)	13.50 (0.45,35.48)	19.90 (11.23,91.78)	49.50 (27.80,137.00)	24.763	<0.001
WBC (10^9/L)	7.30 (5.20,10.10)	6.00 (4.50,7.65)	10.50 (6.55,13.68)	7.30 (3.20,8.30)	9.528	0.023
HCT (%)	34.30 (30.80,38.10)	38.95 (34.80,42.08)	40.05 (35.25,40.88)	37.00 (31.40,40.30)	14.158	0.003
ESR (mm/1h)	36.00 (19.00,65.00)	18.00 (7.25,29.75)	50.00 (30.00,62.75)	31.00 (5.00,47.00)	14.722	0.002
ADA (U/L)	13.70 (10.00,20.60)	16.15 (11.78,26.93)	14.25 (8.28,33.60)	28.30 (15.10,39.00)	13.588	0.004
ALP (U/L)	75.80 (61.6,114.70)	66.65 (54.90,90.03)	68.35 (57.38,120.15)	119.20 (71.30,327.80)	7.590	0.055
ALT (U/L)	20.60 (13.30,45.80)	32.60 (13.65,64.68)	16.25 (11.08,26.58)	56.10 (34.30,108.40)	15.774	0.001
AST (U/L)	22.80 (17.50,39.80)	26.65 (19.30,57.53)	19.15 (17.05,31.38)	61.90 (25.00,93.20)	13.692	0.003
IBIL (μmol/L)	5.20 (3.60,8.00)	6.85 (4.63,9.23)	3.75 (2.85,7.78)	7.60 (4.80,12.70)	9.723	0.021
LDH (mg/dL)	195.80 (171.00,262.20)	254.75 (185.05,416.98)	228.80 (170.05,378.55)	283.00 (197.80,382.50)	14.965	0.002
AG (mmol/L)	13.00 (11.40,14.40)	13.45 (12.40,14.58)	16.05 (12.63,16.73)	13.10 (12.30,15.40)	9.516	0.023
Mg (mmol/L)	0.82 (0.77,0.91)	0.88 (0.83,0.95)	0.89 (0.65,0.99)	0.92 (0.85,1.00)	14.772	0.002
ALB (g/L)	33.30 (29.60,36.30)	36.00 (32.85,38.53)	32.50 (28.93,41.13)	33.60 (29.60,38.50)	10.074	0.018
C3 (g/L)	0.98 (0.85,1.13)	0.98 (0.78,1.10)	0.83 (0.77,0.89)	0.73 (0.70,0.93)	19.289	<0.001
C4 (g/L)	0.28 (0.22,0.34)	0.27 (0.21,0.32)	0.22 (0.13,0.26)	0.25 (0.20,0.30)	11.935	0.008
PCT (ng/mL)	0.14 (0.06,0.30)	0.20 (0.05,0.88)	0.37 (0.05,0.63)	0.23 (0.19,0.34)	3.614	0.306
T3 (pg/mL)	2.12 (1.77,2.57)	2.05 (1.76,2.74)	2.38 (2.00,2.63)	1.91 (1.60,2.03)	7.262	0.064

Table 4 — Evaluation of machine learning (ML) models

	Accuracy	Precision		Recall		F1-score	
		Micro	Weighted	Micro	Weighted	Micro	Weighted
RF	0.7175	0.7292	0.6660	0.6731	0.6731	0.7000	0.6309
XGBoost	0.6549	0.7347	0.6779	0.6923	0.6923	0.7129	0.6618
LightGBM	0.6782	0.7021	0.5680	0.6346	0.6346	0.6667	0.5949

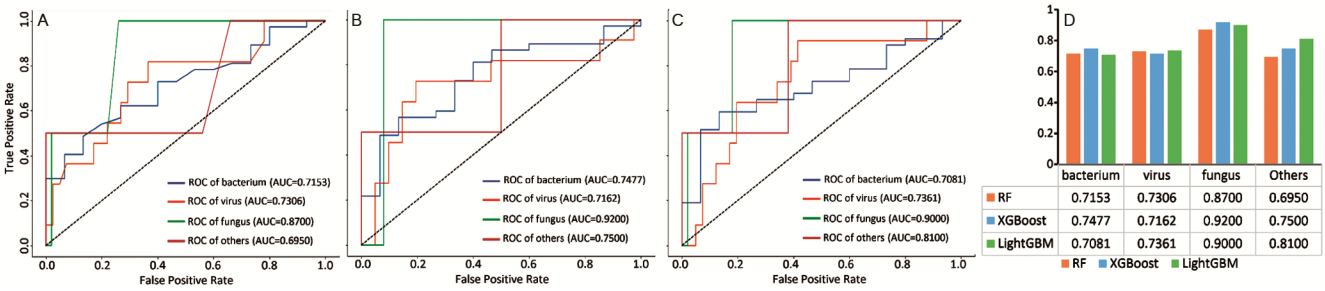

Fig. 1 — AUCs of different etiologies based on machine learning (ML). (A–C) AUCs of RF, XGBoost and LightGBM for different etiologies; and (D) AUCs of MLs shown by bar chart.

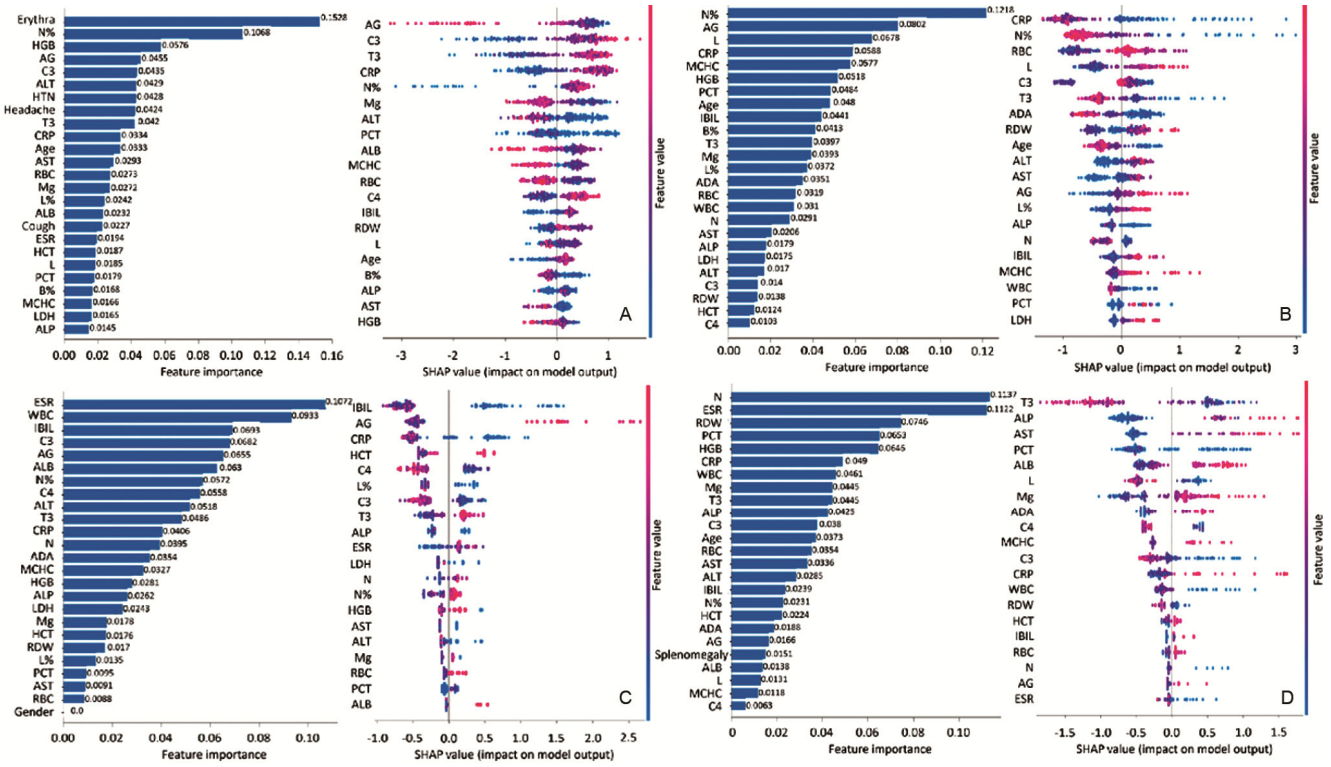

Fig. 2 — Important contribution of characteristics in different infections and their SHAP score. (A) Bacterial IDs; (B) Virus IDs; (C) Fungal IDs; and (D) Other pathogens. [SHAP, The vertical axis ranks features according to the sum of SHAP values of all samples, and horizontal axis is the SHAP value (distribution of features on the model output). Each point represents a sample, and the colour represents the values (red for high values and blue for low values. SHAP values <0 indicate negative effect on model output, whereas SHAP values >0 indicate positive effect on model output]

age were the top 10 contribution indices of bacterial infection, and high SHAP values of C3, T3, CRP, N%, C4, and age were beneficial to the classifier output model. For viral infection, N% was the most important index, followed by AG, L, CRP, MCHC, HGB, PCT, age, IBIL and B%; higher RBC, L, and AG and lower CRP, N%, T3, and age indicated viral infection (Fig. 2B). Meanwhile, AG, T3, ESR, and N% and lower C3 and C4 values were beneficial to identify fungal infection (Fig. 2C). Lower T3, C3, C4 and L were conducive to other pathogens' ID model output (Fig. 2D).

Discussion

The etiology of ID of classic FUO is complex. Clinical drug application depends on the pathogen types. Accurately distinguish types of diseases is crucial. In this study, multiple ML algorithms were used to construct a multiclassification prediction model for the etiology of IDs of classic FUO in adults, and the prediction models were explained. At present, studies on the IDs of classic FUO based on ML are few. Compared with Yan *et al.*[12], the etiological types of IDs of classic FUO were further subdivided in the present study to construct models to improve the accuracy of etiology judgment and clinical medication, and reduce the possibility of drug resistance caused by inaccurate medication.

The results showed that bacterial infections were the main cause of IDs. Analysis of specific pathogens demonstrated that *Mycobacterium tuberculosis* infection accounted for 12.31%, and tuberculosis accounted for 45.83%, both of which decreased[15,16]. This phenomenon may be related to the early diagnosis of pulmonary tuberculosis, drug intervention, and increased attention to NTM-PD diagnosis in China. With the improvement of living standards and the use of preventive antibiotics, the incidence of rheumatic heart disease caused by streptococcal infection has been reduced[17]. However, infective endocarditis (IE) still accounted for 12.31% of bacterial IDs in this study. A number of epidemiological studies have shown that the incidence of IE was increasing in economically developed areas, and its occurrence was related to the increase in invasive procedures[18-20]. The use of long-term

intravenous catheterization, prosthetic heart valves, and permanent pacemakers has greatly increased the incidence of staphylococcal infections leading to IE[21-24]. Therefore, physicians should be alert to the possibility of IE when diagnosing patients with a long history of fever. A study found that the diagnostic rate of *Chlamydia psittaci* pneumonia has increased in the past 3 years may be because the diagnosis of psittacosis has been improved and the delay of diagnosis was reduced by the implementation of metagenomic next-generation seqencing detection in each hospital[25,26].

Analysis of clinical symptoms revealed that rash was the only concomitant symptom with significant differences and contributed significantly to the model output. Bacteria, viruses and others could cause rashes. The shape of the rash varies with different etiologies[27].Rash has a certain importance to the differential diagnosis of fever, which should be paid attention to by clinicians. As a traditional index of inflammation, CRP has not shown prominent importance in the model output, possibly due to its low specificity[28]. High SHAP values of C3, CRP, N% and C4 were positively correlated with the model output of bacterial IDs, as supported by research results[29-32].

When samples in the classification are unbalanced, the ML algorithm tends to be affected by large samples[33]. In this study, the performance of the model was evaluated by applying micro-average and weighted-average to solve this problem. Different integration models (RF, XGBoost and LightGBM), which have the ability to integrate multiple models to achieve better results than a single model, were selected. RF, as an ensemble supervised learning method composed of multiple decision trees corresponding to various sub-datasets[34], has high model accuracy. Each tree calculates the results and obtains the average of the prediction outcomes. This approach allows reducing variance in decision trees. In this study, the micro-F1 and weighted-F1 of RF were next to those of the XGBoost model. XGBoost is an ensemble of multiple classification and regression tree, which is widely used because of its high interpretability and the ability to identify the most important predictor variables[35,36]. The XGBoost algorithm can indicate the contributions of each of the predictors, making it possible to choose the most relevant predictors. In the present study, precision, recall, and F1-score demonstrated the superiority of the XGBoost model's prediction ability. Thus, XGBoost was chosen, and SHAP was used to explain the relationship between the features and the model output, thereby solving the "black box" problem of ML.

Although this study innovatively predicts the etiology of classic FUO from the perspective of ML, some shortcomings still exist in the research process. First, this study was a retrospective study with a small sample size. Due to the lack of an external validation set, overfitting may be present within the data set, reducing the accuracy of the prediction model. Prospective studies or multicenter studies will be conducted to expand the sample size to further verify the model. Second, this study only included common clinical examination indicators, hence not comprehensive, and it will be further improved in the future.

Conclusion

With above retrospective analysis on the infectious diseases (IDs) of classic fever of unknown origin (FUO) screening 25 indices with statistical differences using multiclassification constructed prediction models of 10-fold cross-validation of Machine learning (ML) viz., Random forest (RF), Light gradients boosting (Light GBM) and Extreme gradient boosting (XGBoost), as explained by SHapley Additive exPlanations (SHAP), we can conclude that bacterial infections rather than viral fungal and other pathogens, are the main cause of IDs of classic FUO. Further, the XGBoost has been proved to be an excellent etiological prediction model for IDs of classic FUO. Findings of this study could help clinicians to predict the etiology quickly and accurately, and it has an auxiliary role in the clinical application of reasonable anti-infective drugs.

Acknowledgement

This work was supported by the Public welfare application research project of Zhejiang Science and Technology Department (No. LGF22H190006).

Ethics approval and Informed Consent

This study complied with medical ethics standards and was approved by the Ethics Committee of Huzhou Central Hospital, Huzhou University.

Conflict of Interest

Authors declare no competing interests.

References

1 Erdem H, Baymakova M, Alkan S, Letaief A, Yahia WB, Dayyab F, Kolovani E, Grgic S, Cosentino F, Hasanoglu I, Khedr R, Marino A, Pekok AU, Eser F, Arapovic J, Guner HR, Miftode IL, Poposki K, Sanlidag G, Tahmaz A, Sipahi OR, Miftode EG, Oncu S, Cagla-Sonmezer M, Addepalli SK, Darazam IA, Kumari HP, Koc MM, Kumar MR, Sayana SB, Wegdan AA, Amer F, Ceylan MR, El-Kholy A, Onder T, Tehrani HA, Hakamifard A, Kayaaslan B, Shehata G, Caskurlu H, El-Sayed NM, Mortazavi SE, Pourali M, Elbahr U, Kulzhanova S, Yetisyigit T, Saad SA, Cag Y, Eser-Karlidag G, Pshenichnaya N, Belitova M, Akhtar N, Al-Majid F, Ayhan M, Khan MA, Lanzafame M, Makek MJ, Nsutebu E, Cascio A, Dindar-Demiray EK, Evren EU, Kalas R, Kalem AK, Baljić R, Ikram A, Kaya S, Liskova A, Szabo BG, Rahimi BA, Mutlu-Yilmaz E, Sener A & Rello J, Classical fever of unknown origin in 21 countries with different economic development: an international ID-IRI study. *Eur J Clin Microbiol Infect Dis*, 42 (2023) 387-398.

2 Haidar G & Singh N, Fever of Unknown Origin. *N Engl J Med*, 386 (2022) 463.

3 Wright WF, Mulders-Manders CM, Auwaerter PG & Bleeker-Rovers CP, Fever of Unknown Origin (FUO) - A Call for New Research Standards and Updated Clinical Management. *Am J Med*, 135 (2022) 173.

4 Fusco FM, Pisapia R, Nardiello S, Cicala SD, Gaeta GB & Brancaccio G, Fever of unknown origin (FUO): which are the factors influencing the final diagnosis? A 2005-2015 systematic review. *BMC Infect Dis*, 19 (2019) 653.

5 Van Rijsewijk ND, IJpma FFA, Wouthuyzen-Bakker M & Glaudemans AWJM, Molecular Imaging of Fever of Unknown Origin: An Update. *Semin Nucl Med*, 53 (2023) 4.

6 Dong Y, Gao Y, Chai Y & Shou S, Use of Quantitative Metagenomics Next-Generation Sequencing to Confirm Fever of Unknown Origin and Infectious Disease. *Front Microbiol*, 13 (2022) 931058.

7 Fu ZF, Zhang HC, Zhang Y, Cui P, Zhou Y, Wang HY, Lin K, Zhou X, Wu J, Wu HL, Zhang WH & Ai JW, Evaluations of Clinical Utilization of Metagenomic Next-Generation Sequencing in Adults with Fever of Unknown Origin. *Front Cell Infect Microbiol*, 11 (2022) 745156.

8 Wright WF, Simner PJ, Carroll KC & Auwaerter PG, Progress Report: Next-Generation Sequencing, Multiplex Polymerase Chain Reaction, and Broad-Range Molecular Assays as Diagnostic Tools for Fever of Unknown Origin Investigations in Adults. *Clin Infect Dis*, 74 (2022) 924.

9 David A & Quinlan JD, Fever of Unknown Origin in Adults. *Am Fam Physician*, 105 (2022) 137.

10 Zou Q & Ma Q, The application of machine learning to disease diagnosis and treatment. *Math Biosci*, 320 (2020) 108305.

11 Ngiam KY & Khor IW, Big data and machine learning algorithms for health-care delivery. *Lancet Oncol*, 20 (2019) e262-e273.

12 Yan Y, Chen C, Liu Y, Zhang Z, Xu L & Pu K, Application of Machine Learning for the Prediction of Etiological Types of Classic Fever of Unknown Origin. *Front Public Health*, 9 (2021) 800549.

13 Chen T & Guestrin C, Xgboost: a scalable tree boosting system. In: *Proc 22nd ACM SIGKDD Int Conf Knowledge Discovery and Data Mining*, *San Francisco*, CA, (Association for Computing Machinery, NY, US), 2016, 785.

14 Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q & Liu T, LightGBM: A highly efficient gradient boosting decision tree. In: *Proceedings of the 31st Conference on Neural Information Processing Systems*, (NIPS 2017, Long Beach, CA), 2017.

15 Zhou G, Zhou Y, Zhong C, Ye H, Liu Z, Liu Y, Tang G, Qu J & Lv X, Retrospective analysis of 1,641 cases of classic fever of unknown origin. *Ann Transl Med*, 8 (2020) 690.

16 Shi X, Zhang L, Zhang Y, Zhou B *&* Liu X*,* Utility of T-Cell Interferon-γ Release Assays for Etiological Diagnosis of Classic Fever of Unknown Origin in a High Tuberculosis Endemic Area--a pilot prospective cohort. *PLoS ONE,* 11 (2016) e0146879.

17 Torres RPA, Torres RFA, de Crombrugghe G, Moraes da Silva SP, Cordeiro SLV, Bosi KA, Smeesters PR & Torres RSLA, Improvement of Rheumatic Valvular Heart Disease in Patients Undergoing Prolonged Antibiotic Prophylaxis. *Front Cardiovasc Med*, 8 (2021) 676098.

18 Nappi F & Spadaccio C, keep fumbling around in the dark when it comes to infective endocarditis, or produce new, reliable data to redesign the guidelines? *J Thorac Cardiovasc Surg*, 155 (2018) 75.

19 Tinica G, Tarus A, Enache M, Artene B, Rotaru I, Bacusca A & Burlacu A, Infective endocarditis after TAVI: a meta-analysis and systematic review of epidemiology, risk factors and clinical consequences. *Rev Cardiovasc Med*, 21 (2020) 263.

20 Janszky I, Gémes K, Ahnve S, Asgeirsson H & Möller J, Invasive Procedures Associated with the Development of Infective Endocarditis. *J Am Coll Cardiol,* 71 (2018) 2744.

21 Greenspon AJ, Patel JD, Lau E, Ochoa JA, Frisch DR, Ho RT, Pavri BB & Kurtz SM, 16-year trends in the infection burden for pacemakers and implantable cardioverter-defibrillators in the United States 1993 to 2008. *J Am Coll Cardiol*, 58 (2011) 1001.

22 Han HC, Hawkins NM, Pearman CM, Birnie DH & Krahn AD, Epidemiology of cardiac implantable electronic device infections: incidence and risk factors. *Europace*, 23 (2021) iv3.

23 Sawbridge D, Taylor M, Teubner A, Abraham A, Woolfson P, Abidin N, Chadwick PR & Lal S, Infective Endocarditis in Patients with Intestinal Failure: Experience from a National Referral Center. *J Parenter Enteral Nutr*, 45 (2021) 309.

24 McCarthy JT & Steckelberg JM, Infective endocarditis in patients receiving long-term hemodialysis. *Mayo Clin Proc*, 75 (2000) 1008.

25 Chen X, Cao K, Wei Y, Qian Y, Liang J, Dong D, Tang J, Zhu Z, Gu Q & Yu W, Metagenomic next-generation sequencing in the diagnosis of severe pneumonias caused by Chlamydia psittaci. *Infection*, 48 (2020) 535.

26 Wang K, Liu X, Liu H, Li P, Lin Y, Yin D, Yang L, Li J, Li S, Jia L, Bai C, Jiang Y, Li P & Song H, Metagenomic diagnosis of severe psittacosis using multiple sequencing platforms. *BMC Genomics*, 22 (2021) 406.

27 Goldman L & Schafer AI, *Goldman's Cecil Medicine*, 26th ed., (Elsevier Inc., Amsterdam, Netherlands), 2019.

28 Sproston NR & Ashworth JJ, Role of C-Reactive Protein at Sites of Inflammation and Infection. *Front Immunol*, 9 (2018) 754.

29 Wang H & Liu M, Complement C4, Infections, and Autoimmune Diseases. *Front Immunol*, 12 (2021) 694928.
30 Xu T, Wang L, Wu S, Zhou F & Huang H, Utility of a Simple Scoring System in Differentiating Bacterial Infections in Cases of Fever of Unknown Origin. *Clin Infect Dis*, 71 (2020) S409.
31 Corcoran JA & Napier BA, C3aR plays both sides in regulating resistance to bacterial infections. *PLoS Pathog*, 18 (2022) e1010657.
32 Li Y, Min L & Zhang X, Usefulness of procalcitonin (PCT), C-reactive protein (CRP), and white blood cell (WBC) levels in the differential diagnosis of acute bacterial, viral, and mycoplasmal respiratory tract infections in children. *BMC Pulm Med*, 21 (2021) 386.
33 Lin WJ & Chen JJ, Class-imbalanced classifiers for high-dimensional data. *Brief Bioinform*, 14 (2013) 13-26.
34 Breiman, L, Manual for Setting Up, Using and Understanding Random Forest V40. Available online:https://www.stat. berkeley. edu/~breiman/Using_random_forests_v4.0.pdf.
35 Rhys HI, *Machine Learning with R, the tidyverse and mlr*. (Manning Publications Co. NY, USA), 2020.
36 Singh KP & Gupta S, *In silico* prediction of toxicity of non-congeneric industrial chemicals using ensemble learning based modeling approaches. *Toxicol Appl Pharmacol*, 275 (2014) 198212.