Construction of a Prevalence Prediction Model for COVID-19 in Counties of Guizhou Province: Based on the Random Forest Algorithm

Shouyan Chen[1], Li Xi[1], Junhua Wang[1], Jiangping Zhang[1,2,*]

1School of Public Health, Key Laboratory of Environmental Pollution Monitoring and Disease Control, Ministry of Education, Guizhou Medical University, Guiyang, Guizhou, China

2Guiyang Public Health Treatment Center, Guiyang, Guizhou, China

**Corresponding Author*

Abstract: This research intends to construct a county-level prediction model for COVID-19 prevalence in Guizhou Province based on the Random Forest algorithm. The purpose was to offer a foundationfor regional epidemic risk early-warning and the optimal allocation of prevention and control resources. From February to March 2020, a survey on COVID-19 protection knowledge was carried out among 1,987 residents in Guizhou Province through "Wenjuanxing" platform. Data regarding demographic characteristics and eight knowledge dimensions were gathered. The Random Forest algorithm was utilized for feature selection and model building. Optimal parameters were ascertained via 5-fold cross-validation. SHAP values were utilized for feature interpretation, and the model's performance was further compared with XGBoost and Linear Regression.On the test dataset, the Random Forest model yielded an MSE of 0.0930, an RMSE of 0.3049, and an MAE of 0.1841. These results indicated that it had better predictive performance comparedto the reference models. The five most important features were: Knowledge of mental health intervention (Q8.3), Total knowledge score (point), Basic knowledge score (Q1), Knowledge of transmission routes (Q1.1), and Knowledge of community prevention and control (Q7). SHAP analysis further confirmed the positive impacts of thesefeatures on the prediction results. The prediction model constructed with the RandomForest algorithm shows high accuracy and interpretability. It can pinpoint the key factors influencing the risk of COVID-19 prevalence in Guizhou's counties, thereby providing scientific backing for formulating differentiated prevention and control strategies.

Keywords: COVID-19; Prediction of Prevalence; Random Forest; Machine Learning

1. Introduction

In late 2019, the COVID-19 outbreak spread at a rapid pace and caused widespread social impacts across the globe. To tackle the health crisis, the Chinese government swiftly implemented rigorous prevention and control policies, effectively restraining the spread of the coronavirus. At this stage, the whole country has stepped into steady normalized epidemic management [1,2]. In this stage, continuously tracking the dynamic changes of the epidemic and accurately predicting the transmission risks in various regions are crucial for the efficient allocation of limited prevention and control resources and the improvement of emergency response efficiency [3]. Guizhou Province is located in the southwest of China. Its unique karst landform leads to a scattered population distribution, endowing the transmission pattern of the epidemic with distinctive regional characteristics [4]. Data from 2022 showed that even under the same prevention and control policies, there were still significant differences in the COVID-19 prevalence among different districts and counties within Guizhou Province. Further research revealed that in addition to objective factors such as geographical conditions and economic levels, the uneven levels of public COVID-19 prevention knowledge and compliance with health behaviors across regions were one of the important causes of such prevalence differences [5]. Studies have demonstrated that scientific and systematic prevention knowledge can guide individuals to develop good protective habits, thereby reducing

the risk of infection. In areas with higher levels of prevention knowledge, the risk of COVID-19 infection can be reduced by approximately 23.5% [6]. Therefore,against the backdrop of normalized epidemic prevention and control, constructing a prediction model integrating regional population characteristics and public prevention knowledge levels is of great practical significance for realizing early warning of regional epidemic risks and promoting the precise allocation of prevention and control resources [7]. The proposed model can rapidly detect high-risk zones of the epidemic and lay a scientific foundation for adopting targeted prevention and control plans [8]. With the extensive application of artificial intelligence technology in the medical field, the random forest algorithm, which can effectively identify nonlinear relationships and interactions among variables, has been increasingly adopted in studies on disease risk prediction [9,10]. To conclude, this research develops a regional COVID-19 risk prediction model for Guizhou Province, taking local population features and residents ' public health awareness into comprehensive consideration. Its core objectives include accurately pinpointing high-risk locations, optimizing the efficient allocation of pandemic prevention and control resources, and furnishing scientific evidence for localized, targeted containment policies.

2. Subjects and Methods

2.1 Study Subjects

This study was conducted from February 19, 2020 to March 4, 2020. Anonymous online questionnaires were randomly distributed via the "Wenjuanxing" platform, targeting various institutions in Guizhou Province, including enterprises, public institutions, universities and communities. Investigators received unified training before distributing the online questionnaires and providing guidance for respondents. A total of 2,059 questionnaires were collected. After excluding 72 invalid responses, 1,987 valid questionnaires were ultimately obtained, with an effective recovery rate of 96.5%. The present research has obtained ethical approval from the Ethics Committee of Guiyang Center for Disease Control and Prevention, and informed consent was obtained from all participants.

2.2 Survey Methods

The questionnaire consisted of 8 cognitive modules (40 items) and a section on demographic information. The cognitive modules covered: basic knowledge (5 items), personal protection (7 items), protection for the elderly (5 items), protection for children (4 items), protection for pregnant and lying-in women (3 items), home disinfection (4 items), community prevention and control (5 items), and mental health intervention (4 items). This survey instrument exhibited sound psychometric properties, and its internal consistency coefficient (Cronbach's α) was calculated as 0.80. All items in the questionnaire were single-choice questions. Respondents were required to complete their personal information before submission; correct answers were displayed only after the questionnaire was submitted. To ensure the authenticity of the survey, consistent terminology was used throughout the questionnaire, and no private information such as respondents' names, telephone numbers, or ID numbers was collected. Each IP address was restricted to one response, the completion time was set to no less than 100 seconds, and the questionnaire could only be submitted when fully completed.

2.3 Model Construction and Evaluation

Feature selection was first conducted by adopting the random forest algorithm to screen variable indicators, thereby supporting the subsequent development of the prediction model. In terms of data partitioning, the original dataset was split into a training set and a test set at an 8:2 ratio, ensuring that both subsets possessed adequate representativeness. For model training, the R language was applied to fit the random forest model, with the gradient boosting approach adopted to optimize its core parameters. Besides, grid search was utilized to adjust hyperparameters of the XGBoost model, and stepwise regression was performed to filter out influential and significant variables. Finally, model evaluation was carried out in the follow-up research. The test set was used as the internal validation dataset, and mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) were selected as the core evaluation metrics to reflect the prediction accuracy of the model. Model interpretation and visualization. To clarify the influence of each predictor variable on the model output, this

study performed model interpretation analysis based on SHAP [11]. Meanwhile, visualization methods were applied to present the ranking of variable importance and the influence mechanism, enhancing the comprehensibility and persuasiveness of the model results.

3. Results

3.1 General Sociodemographic Characteristics

A total of 1987 participants were enrolled in this study. The median regional COVID-19 prevalence (per 100,000 population) was 0.3785. There were 436 males (21.94%) and 1551 females (78.06%), No obvious statistical discrepancy existed across different groups (P=0.098). By age group: 1171 participants aged 16-35 years (58.93%), 696 aged 36-55 years (35.03%), and 120 aged over 56 years (6.04%), with significant differences observed across groups(P<0.001). By occupation: 883 employed individuals (44.44%), 314 students (15.80%), and 790 with other occupations (39.76%), with statistically significant differences among groups (P<0.001).By cognitive level: 487 participants with passing or below scores (24.5%), 539 with average level (27.12%), 509with good level (25.62%), and 452 with excellent level (22.75%), with no significant variations observed across groups (P=0.999). See Table 1.

Table 1. Analysis of Basic Demographic Characteristics

various	N (%)	Mean ($\times10^5$)	Median ($\times10^5$)	Mean rank	*P*
N	1987 (100)	0.4372	0.3785		
gender					
male	436 (21.94)	0.4229	0.3785	962.27	0.098
female	1551 (78.06)	0.4412	0.3785	1002.92	
age					
16–35 years old	1171 (58.93)	0.4191	0.3785	953.66	<0.001
36–55 years old	696 (35.03)	0.4648	0.3785	1051.42	
≥56 years old	120 (6.04)	0.4532	0.3785	1054.57	
Occupation					
Employed	883(44.44)	0.4564	0.3785	1055.28	<0.001
Student	314(15.80)	0.3364	0.1526	699.76	
Others	790(39.76)	0.4557	0.3785	1042.45	
Cognitive Level					
Below average or failing	487(24.5)	0.4297	0.3785	991.89	0.999
Average	539(27.12)	0.4448	0.3785	996.42	
Good	509(25.62)	0.4335	0.3785	994.16	
Excellent	452(22.75)	0.4404	0.3785	993.21	

Notes: Cognitive level was classified based on the quartile range of the total questionnaire scores. Scores below the 25th percentile (P_{25}) indicated pass or below cognitive level; scores between P_{25} and P_{50} indicated moderate cognitive level; scores between P_{50} and P_{75} indicated good cognitive level; and scores above P_{75} indicated excellent cognitive level. Prevalence was calculated using the real-time number of confirmed cases on the day of questionnaire completion in the respondents' reported residence (district/county level) divided by the person-years of the local population.

3.2 Construction of a COVID-19 Prevalence Prediction Model for Various Regions in Guizhou

In total, 1,987 samples were randomly split into a training set (n=1,590) and a test set (n=397) at an 8:2 ratio to establish and verify a random forest prediction model.

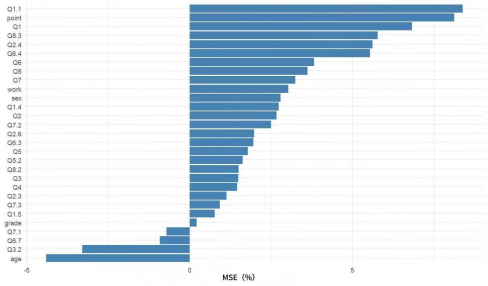
3.2.1 Feature selection and importance ranking

In this study, the random forest algorithm was used to evaluate the feature importance of variables across various knowledge dimensions and demographic variables included in the questionnaire. In Figure 1a (without noise features), Q1.1 (What are the main transmission routes of COVID-19?), Q1 (score of basic knowledge section), and Q8.3 (What psychological adjustment methods are available for frontline medical staff during the pandemic?) were the top-ranked important features. In the importance ranking results incorporating noise features (Figure 1b), the top five features in descending order of importance were: Q8.3, point (total knowledge score), Q1, Q1.1, and Q7 (score of community prevention and control section). Notably, the three noise features (noise1, noise2, noise3) ranked last in the importance list. This result indicates that the constructed model can effectively distinguish real features from noise features and has good ability to identify practically meaningful real features, further verifying the reliability of the feature selection results.

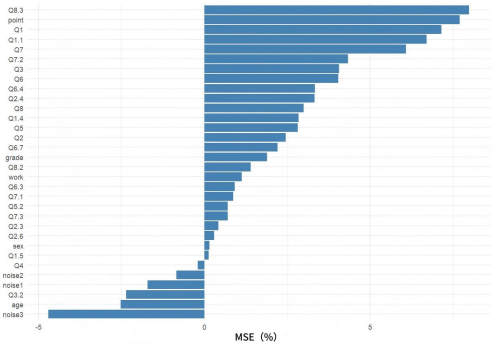
3.2.2 Model construction and parameter optimization

In the present study, 5-fold cross-validation was employed to determine the optimal parameters for the random forest model. According to the

results, the optimal mtry value (number of randomly selected variables at each node split) was 2. Meanwhile, the optimal number of decision trees (ntree) was determined based on out-of-bag (OOB) error analysis: when ntree was less than 600, the OOB error decreased significantly with an increasing number of decision trees; when ntree reached 600, the OOB error dropped to 0.092 and tended to stabilize. A further increase in the number of decision trees produced no obvious change in the OOB error but instead raised the computational cost of the model. Ultimately, ntree = 600 was determined as the optimal number of decision trees for the random forest model. This parameter configuration ensures satisfactory model fitting performance while effectively mitigating overfitting (see Figure 2).



a). Feature Importance Ranking Based on MSE Increment



b). Importance Ranking with Noise Features, Where Noise Variables Rank Las

Figure 1. Ranking of Feature Importance

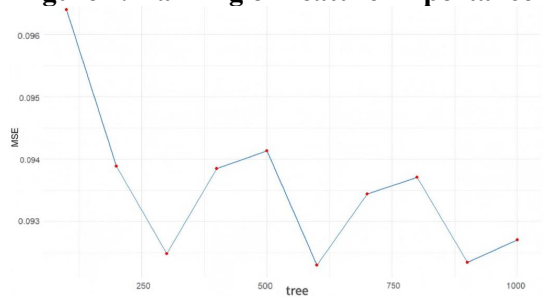


Figure 2. Curve of Random Forest MSE with the Number of Trees, Showing That the Error Tends to Stabilize at Approximately 600 Trees

3.2.3 Model interpretation and shap analysis

In this study, SHAP values were used to visually interpret the prediction outputs of the random forest model (Figure 3). In the figure, the color intensity of the features corresponds to their value levels: a darker color indicates a higher feature value. Taking Q8.3 (knowledge of mental health interventions) as an example, when the score of this feature is high (dark-colored regions in the figure), the corresponding SHAP values tend to be positive, indicating that the feature makes a more prominent contribution to predicting a high prevalence rate. Conversely, when the feature value is low (light-colored regions), its SHAP values tend to be negative, contributing more significantly to predicting a low prevalence rate. Overall, Q8.3 ("What psychological adjustment methods are available for frontline medical staff during the pandemic?"), point (total knowledge score), and Q1 (score of basic knowledge section) are the three most influential features on the model prediction results. Higher SHAP values corresponding to the features indicate a stronger positive effect on predicting a high epidemic risk.

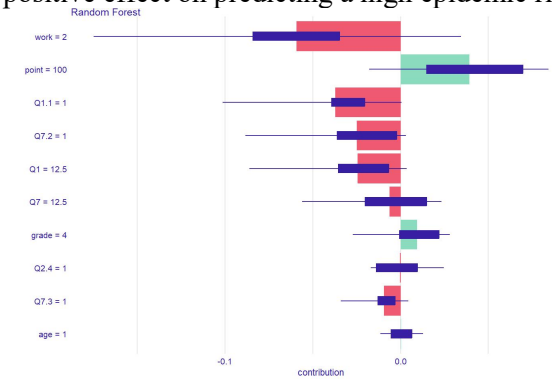


Figure 3. Model Interpretation Plot Based on SHAP Values

3.3 Model Performance Evaluation, Comparison and Robustness Verification

Validation was conducted on the test set, and the model demonstrated satisfactory predictive performance, with an MSE of 0.0930, RMSE of 0.3050, and MAE of 0.1841 (Table 2).The model was further compared with the XGBoost model and linear regression model in terms of performance (see Figure 4). The results indicated that the random forest model performed better than the other two models in terms of RMSE, MAE and other indicators. After adding noise features, the MSE of the random forest model decreased from 0.0930 to 0.0850. This may be attributed to a slight reduction effect of random fluctuations in the noise features on the model

prediction error. The above results indicate that the proposed random forest model has strong robustness and can maintain stable prediction performance under complex data conditions.

Table 2. Model Performance Evaluation Metrics

Indicator	Accuracy
MSE	0.0930
RMSE	0.3049
MAE	0.1841

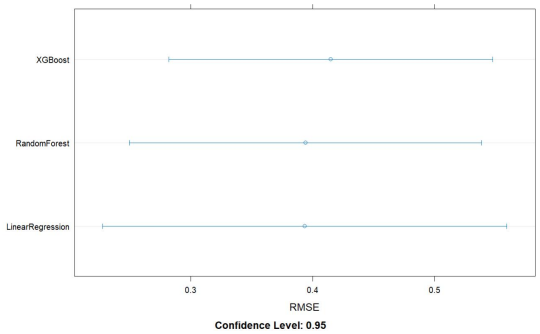


Figure 4. Comparison Plot of RMSE among Models

4. Discussion

In this research, a county-level random forest model was developed to predict COVID-19 prevalence, using demographic characteristics and COVID-19 prevention knowledge as key variables. The model was verified to have favorable prediction performance, with MSE = 0.0930, RMSE = 0.3049, and MAE = 0.1841. SHAP analysis was used to clarify the contribution magnitude and directional effect of each feature variable to the model prediction results, providing a theoretical basis for targeted prevention and control in COVID-19 risk early warning under the normalized epidemic management in Guizhou Province.

4.1 Model Performance and Feature Interpretation

Numerous existing studies have demonstrated that random forest exhibits certain advantages in constructing COVID-19-related prediction models. Studies have shown that the random forest model provides better prediction performance than linear models for the cumulative number of COVID-19 cases in the following two weeks [12]. Ren [13] developed a three-step prediction model (TSPM-ML), which combined algorithms such as random forest and seq2seq with surveillance data to forecast the actual infection scale and confirmed case numbers of the epidemic. The predicted epidemic outcomes for six countries were largely consistent with real-world data, confirming the practical application value of this model in clinical settings. In addition, the importance of total knowledge score and mental health intervention in this study is consistent with the findings of Deng [14]. Their team confirmed through structural equation modeling that health literacy is significantly associated with COVID-19 prevention knowledge and behaviors, and that public knowledge of protective measures plays a key role in epidemic-related prediction. Other studies have shown that a prediction accuracy of 97% can still be achieved even when models are built using only partial features, further supporting the excellent robustness of this algorithm for high-dimensional data [15].

4.2 Relationship between Knowledge Level and Epidemic Prevention and Control

This study found no significant differences among groups with different cognitive levels. However, the importance of total knowledge score for prediction results revealed by SHAP analysis is consistent with the conclusions of previous studies [16,17]. A survey by Zhang [18] on COVID-19 vaccine-related knowledge, attitudes and practices among residents in East China showed that factors such as e-health literacy and occupation significantly affected vaccine-related cognitive levels, indirectly indicating that public knowledge of protective measures can influence epidemic transmission by regulating individual health behaviors. Jia [19] conducted a survey among different populations in Shaanxi Province and found that the public awareness rates of core knowledge regarding COVID-19 transmission routes and susceptible populations were only 41.01% and 47.80%, respectively. Moreover, non-medical students and rural residents exhibited even lower cognitive levels. This finding suggests that improving public knowledge of epidemic prevention may enhance the effectiveness of epidemic control by boosting early symptom identification ability and compliance with protective behaviors. In addition, Lu [20] identified demographic characteristics and vaccine-related cognitive levels as key variables predicting vaccination willingness, which indirectly supports the important role of knowledge level in epidemic prevention and control. In this study, knowledge of mental

health intervention (Q8.3) emerged as a strong predictor, confirming that frontline medical workers, as the core force in epidemic prevention and control, have a psychological state directly related to the efficiency and service quality of prevention and control efforts during the pandemic. Medical staff with strong psychological adjustment abilities can reduce work errors caused by psychological stress, thereby lowering the risk of virus transmission [21]. Meanwhile, popularizing epidemic-related psychological intervention knowledge among the public can improve their psychological adaptability to the pandemic and reduce irrational behaviors caused by panic, thereby indirectly lowering the COVID-19 prevalence rate [22]. Therefore, it is recommended that in epidemic prevention and control, training on mental health intervention knowledge for medical staff should be incorporated into routine prevention and control measures. At the same time, communication channels such as community bulletin boards, short-video platforms, and WeChat official accounts should be fully utilized to popularize routine psychological adjustment methods for epidemic management among the public, so as to further strengthen public awareness of mental health protection.

4.3 Practical Application Value of the Model

The model constructed in this study exhibits favorable robustness, consistent with numerous relevant studies [23]. Galasso developed a county-level COVID-19 prediction model that maintained high predictive performance relying only on test positivity data, demonstrating the good applicability of the random forest algorithm even with limited data types [24]; After comparing various machine learning algorithms, compared multiple machine learning algorithms and reported that the random forest model achieved superior performance in forecasting daily new COVID-19 cases [25]. Furthermore, Luo [26] developed a geographically weighted random forest model to investigate the nonlinear association between 47 influencing factors and COVID-19 mortality, providing strong implications for the allocation of regional prevention and control resources, which is consistent with the objectives of this study. The COVID-19 vulnerability index model developed by Tiwari [16] also demonstrated the guiding value of random forest for targeted intervention, supporting the application potential of this study in county-level epidemic prevention and control. In terms of policy application, this model can provide insights for COVID-19 prevention and control decision-making in Guizhou Province through the following approaches. First, counties and districts can be divided into high-, medium-, and low-risk levels, with differentiated intervention measures implemented for regions at different risk tiers. Second, based on the core predictive factors identified by this model, such as knowledge of psychological adjustment during the pandemic, health lectures can be organized and popular science manuals distributed to improve the public's basic level of prevention knowledge.

5. Limitations and Prospects

Several limitations exist in the present study and require further optimization in future research. First, the data collected in this study were from 2020, which differ greatly from the current epidemic situation and virus variation. This may lead to changes in the prediction performance of the model, which needs to be updated with real-time data. Second, females and young and middle-aged populations accounted for a relatively high proportion in the study sample, which may result in insufficient sample representativeness and selection bias, potentially affecting the generalizability of the model prediction results. Third, environmental factors such as climate, transportation and economic level were not considered in this study, which may affect the comprehensiveness of model prediction. In the future, various types of data can be integrated to construct a more complete prediction index system.

Acknowledgments

This paper is supported by Science and Technology Foundation Project of the Health Commission of Guizhou Province (gzwjkj2020-1-181).

References

[1] HUANG C, WANG Y, LI X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet, 2020, 395(10223): 497-506.

[2] WANG C, HORBY P W, HAYDEN F G, et al. A novel coronavirus outbreak of global health concern. Lancet, 2020, 395(10223): 470-473.

[3] WANG QIAN. Spatiotemporal Analysis and Transmission Risk Prediction Model of COVID-19 in China. Yunnan Normal University, 2022.
[4] PING K, LEI M, GOU Y, et al. Epidemiologic Characteristics of COVID-19 in Guizhou Province, China. J Infect Dev Ctries, 2021, 15(3): 389-397.
[5] CHU XIANGYUAN, ZHANG RONG, YUAN PING, et al. Investigation and Analysis of KAP and Satisfaction on COVID-19 Prevention and Control among Urban and Rural Residents in Zunyi City. Journal of Jianghan University (Natural Science Edition), 2023, 51(02): 44-50.
[6] GELDSETZER P. Knowledge and Perceptions of COVID-19 among the General Public in the United States and the United Kingdom: A Cross-sectional Online Survey. Ann Intern Med, 2020, 173(2): 157-160.
[7] LI R, PEI S, CHEN B, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science, 2020, 368(6490): 489-493.
[8] LI Q, GUAN X, WU P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med, 2020, 382(13): 1199-1207.
[9] MORGAN R D, YOUSSI B W, CACAO R, et al. Random Forest Prognostication of Survival and 6-Month Outcome in Pediatric Patients Following Decompressive Craniectomy for Traumatic Brain Injury. World Neurosurg, 2025, 193: 861-867.
[10] FENG QING, HE PEIFENG. Construction of a Risk Prediction Model for Carbapenem-Resistant Acinetobacter baumannii Ventilator- Associated Pneumonia Based on Random Forest Model. Chinese Nursing Research, 2024, 38(19): 3410-3416.
[11] LE T, KIM H, KANG H, et al. Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method. Sensors (Basel), 2022, 22(3).
[12] DEJUS B, CACIVKINS P, GUDRA D, et al. Wastewater-based prediction of COVID-19 cases using a random forest algorithm with strain prevalence data: A case study of five municipalities in Latvia. Sci Total Environ, 2023, 891: 164519.
[13] REN JIANQIANG, CUI YAPENG, NI SHUNJIANG. COVID-19 Epidemic Trend Prediction Method Based on Machine Learning. Journal of Tsinghua University (Science and Technology), 2023, 63(06): 1003-1011.
[14] DENG LIFANG. A Study on the Relationship between Residents' Health Literacy and COVID-19 Prevention Knowledge and Behavior Based on Structural Equation Modeling. Nanchang University, 2023.
[15] CHEN Y, OUYANG L, BAO F S, et al. A Multimodality Machine Learning Approach to Differentiate Severe and Nonsevere COVID-19: Model Development and Validation. J Med Internet Res, 2021, 23(4): e23948.
[16] TIWARI A, DADHANIA A V, RAGUNATHRAO V A B, et al. Using machine learning to develop a novel COVID-19 Vulnerability Index (C19VI). Sci Total Environ, 2021, 773: 145650.
[17] ZHANG X, MAGGIONI V, HOUSER P, et al. The impact of weather condition and social activity on COVID-19 transmission in the United States. J Environ Manage, 2022, 302 (Pt B): 114085.
[18] ZHANG MING, DU SHAOLING, HUANG LONG, et al. Investigation and Analysis of COVID-19 Vaccine Knowledge, Attitude and Practice among Residents in East China Based on Network Survey. Journal of Jinzhou Medical University (Social Science Edition), 2021, 19(04): 15-19.
[19] JIA LILI, OUYANG JING, QIAO YOULIN, et al. Investigation on COVID-19 Related Knowledge Cognition of Different Populations in Shaanxi Province. Medicine and Society, 2021, 34(02): 94-98.
[20] LU MINGXIA, LI ZHIYUAN, BAI YIRAN, et al. Analysis of COVID-19 Vaccine Cognitive Level, Willingness to Receive Second Booster Immunization and Its Influencing Factors Among 12,117 People Based on New Media. Modern Disease Prevention and Control, 2024, 35(01): 6-10.
[21] ZHENG TAO, HE XINRAN, YUAN HONGXU, et al. Investigation on cognitive level of COVID-19 among medical staff in Anhui Province. Chinese Journal of Disease Control & Prevention, 2021, 25(04): 427-431.
[22] CHEN TAO, YANG YANBING, CHANG YUWEN, et al. Investigation on Mental Health Status and COVID-19 Related

Knowledge and Behavior Among the Elderly in Shanghai Under Normalized Epidemic Prevention and Control. South China Journal of Preventive Medicine, 2022, 48(02): 247-250.

[23]XIONG QIAN, HE LIN, LI XIAOYAN. Investigation on Knowledge, Attitude and Practice of COVID-19 Among the Public in Deyang City and Analysis of Related Factors. Western Journal of Traditional Chinese Medicine, 2021, 34(02): 1-5.

[24]GALASSO J, CAO D M, HOCHBERG R. A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. Chaos Solitons Fractals, 2022, 156: 111779.

[25]ZHENG H, AN S, QIAO B, et al. A data-driven interpretable ensemble framework based on tree models for forecasting the occurrence of COVID-19 in the USA. Environ Sci Pollut Res Int, 2023, 30(5): 13648-13659.

[26]LUO Y, YAN J, MCCLURE S. Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis. Environ Sci Pollut Res Int, 2021, 28(6): 6587-6599.

