BMC Pulmonary Medicine

RESEARCH

Open Access

Forecasting the trend of tuberculosis incidence in Anhui Province based on machine learning optimization algorithm, 2013–2023

Yan Zhang[1†], Huan Ma[2†], Hua Wang[3†], Qing Xia[1], Shasha Wu[1], Jing Meng[1], Panpan Zhu[1], Zhilong Guo[1] and Jing Hou[1*]

Abstract

Tuberculosis has been one of the most common communicable diseases raising global concerns. Accurately predicting the incidence of Tuberculosis remains challenging. Here we constructed a time-series analysis and fusion tool using multi-source data, and aimed to more accurately predict the incidence trend of tuberculosis of Anhui Province from 2013 to 2023. Random forest algorithm (RF), Feature Recursive Elimination (RFE) and Least absolute shrinkage and selection operator (LASSO) were implemented to improve the derivation of features related to infectious diseases and feature work. Based on the characteristics of infectious disease data, a model of RF-RFE-LASSO integrated particle swarm optimization multiple inputs long short term memory recurrent neural network (RRL-PSO-MiLSTM) was created to perform more accurate prediction. Results showed that the PSO-MiLSTM achieved excellent prediction results compared with common single-input and multi-input time-series models (test set MSE:42.3555, MAE: 59.3333, RMSE: 146.7237, MAPE: 2.1133, R^2: 0.8634). PSO-MiLSTM enriches and complements the methodological research content of calibrating the time-series predictive analysis of infectious diseases using multi-source data, and can be used as a brand-new benchmark for the analysis of influencing factors and trend prediction of infectious diseases at the public health level in the future, as well as providing a reference for incidence rate prediction of infectious diseases.

Keywords Time-series analysis, Tuberculosis, Feature engineering, Neural network, Swarm intelligence algorithms

Introduction

Tuberculosis (TB), an infectious disease of the lungs caused by Mycobacterium tuberculosis, is a common infectious disease worldwide. According to the World Health Organization (WHO), the global incidence of TB decreased by 1.5% between 2000 and 2015, but 200 million people are still infected with Mycobacterium tuberculosis, and 1.5 million people die from TB every year [1]. Mathematical models or artificial intelligence models can effectively predict the future epidemiological trends of infectious diseases and develop appropriate prevention and control measures based on the results [2, 3]. Time series methods are often used in previous studies to predict the incidence of infectious diseases. Among them, linear methods are represented by seasonal autoregressive integrated moving average (SARIMA) [4], and nonlinear methods are represented by non-autoregressive (NAR) [5], long short-term

[†]Yan Zhang, Huan Ma and Hua Wang contributed equally to this work.

*Correspondence:
Jing Hou
ahxkhyhoujing@126.com
[1] Third Department of Tuberculosis, Anhui Chest Hospital, 397 Jixi Road, Shushan, Hefei 230000, China
[2] Department of Oncology, The 901th Hospital of Joint Logistic Support Force PLA, Hefei 230032, China
[3] Eight Department of Tuberculosis, Anhui Chest Hospital, Hefei, China

memory (LSTM) [6], and recurrent neural network (RNN) [7], recurrent neural network (long short term memory, LSTM) [6] and nonlinear autoregressive with exogenous input (NARX) [7]. These models usually utilize historical incidence data to make predictions of future incidence and involve the developmental patterns of infectious diseases themselves.

However, the factors affecting the incidence of infectious diseases are complex, and their correlation characteristics involve multiple aspects. Usually, the outbreak of an infectious disease will lead to an increase in news media coverage of the type [8], fluctuations in the stocks of companies or sectors associated with it [9], and even some changes in the socio-economic structure [10]. Studies have shown that changes in climate and pollutant indicators also have a complex impact on the development of infectious diseases [11, 12]. With the development of satellite remote sensing and Internet technologies, it has become possible to access and utilize multi-source data related to infectious diseases. The effective utilization of these data has a high application value and helps to improve the effectiveness of early warning research on infectious diseases. Currently, some previous studies have successfully applied Internet data in the prediction of infectious diseases and revealed the precedence and relevance of Internet data in the prediction of infectious diseases [13, 14]. Therefore, integrated utilization of multi-source data can significantly improve the effectiveness and accuracy of prediction models.

Feature engineering is a necessary step to ensure the accuracy of machine learning methods, and feature derivation, feature extraction, and feature screening can effectively solve the problem of mismatch on the time scale of variables. However, feature extraction categories are complex and feature screening methods are diverse, and choosing appropriate methods to screen out independent variables with large influence factors is one of the difficulties in establishing efficient prediction models [15, 16]. On the one hand, in order to make full use of the data, in addition to the commonly used statistical features (mean, extreme deviation, etc.), the features in the time domain and frequency domain of the time series data are also extracted [17, 18]; on the other hand, for the characteristics of the derived features with high dimensionality, the feature screening study is carried out. Recursive feature elimination method eliminates weak features by repeatedly training the model to avoid overfitting problem and improve the generalization ability of the model [19]. Random forest and LASSO algorithms also have unique advantages in feature screening [20, 21]. By comparing the performance of the algorithms and targeting the characteristics of the variables, we adopt the fusion of the above three methods for screening in order to determine the input variables of the model, aiming to improve the accuracy of the model prediction.

In China, tuberculosis belongs to one of the higher categories of statutory infectious diseases and has obvious time series characteristics [22]. Therefore, based on the TB dataset in Anhui Province, the data structure, input method, feature engineering and parameter optimization of the incidence prediction model were explored for validity and versatility, and innovations were made in the following aspects: (1) Multi-source data can be fully explored for the factors related to the development of infectious diseases, to correct the prediction results of infectious disease data and to improve the model's prediction ability for the perturbation part. (2) By performing comprehensive feature derivation and adopting multiple methods for feature screening, consistency on the time scale of variables is accomplished while achieving full utilization of data. (3) The constructed PSO-MiLSTM model obtains better disease prediction results, which can provide a reference basis for the prediction of the incidence of various infectious diseases.

Methods and materials

Figure 1 shows the technical route of the data and methods in this study.

Data Sourcing and Cleaning

The data on the number of monthly incidence of tuberculosis in Anhui Province were obtained from the Chinese Center for Disease Control and Prevention (CCDC), the Internet search index was obtained from the Baidu search engine, the data on atmospheric pollutants were obtained from the official website of the China Environmental Monitoring Station (CEMS), the data on climate were obtained from the National Meteorological Information Center (NMIC), and the data on stock fluctuations were obtained from the Flush Data Center. The overall time span of the data is from December 2013 to October 2023, a total of 119 months, covering 35 initial variables, mainly through the database for keyword co-occurrence analysis, expert consensus and experience to determine the keyword information of high relevance to the infectious diseases to be analyzed. Stock information is selected primarily from local comprehensive sector indices. Atmospheric pollutant data and climate data are from NEPA and Meteorological Agency respectively, incorporating the specific information of the data as shown in Table 1.

All the research data have been reasonably collected,

Among them, pollutant information and search index have some missing data, and the percentage of

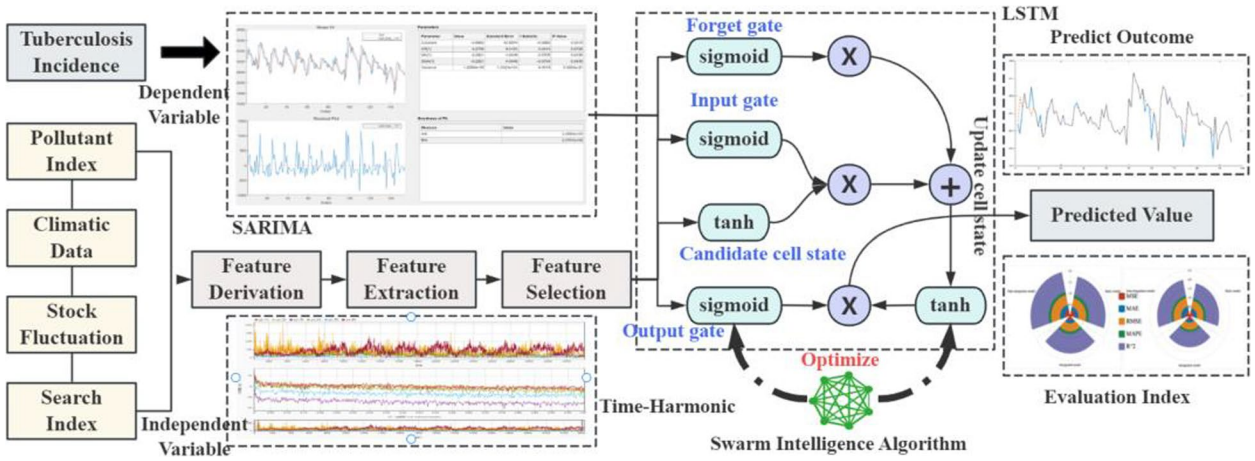

Fig. 1 Full text technology roadmap

Table 1 General information of the experimental datasets

Dataset	Data collectors	Time scale
Tuberculosis statistics data[a]	China Center for Disease Control and Prevention	Month
Search index data[b]	Baidu search engine	Day
Air contaminant data[c]	China National Environmental Monitoring Centre	Day
Climatic data[d]	China Meteorological Administration	3 h
Stock data[e]	Flush data center	Day

[a] https://www.phsciencedata.cn/Share/index.jsp

[b] http://index.baidu.com/v2/index.html

[c] https://www.cnemc.cn/?pc_hash=HQeakg

[d] https://data.cma.cn/

[e] https://www.10jqka.com.cn/

their missing is less than 0.1%, which meets the inclusion criteria, and the simulation filling comparison study was carried out by using mean filling, median filling, proximity filling, and linear interpolation, and it was finally concluded that using the proximity of the backward filling has the smallest impact on the sequence characteristics, so the proximity filling was used to fill the missing data of the pollutant and search index sequences in data cleaning [23], and cross-validation was used to further explore the stability of the data after filling. Therefore, proximity filling is used in data cleaning to fill in the missing data of pollutant and search index series, and cross validation is used to further explore the stability of the filled data. When matching the data, in order to achieve the purpose of early prediction, the independent variable is the previous 1 month data corresponding to the dependent variable. For example, when predicting the number of tuberculosis cases in October 2023, the input data of the independent variable is actually the index of Baidu in September 2023 and other indicators.

Feature extraction and screening

As can be seen from Table 1 and Table 2, there is a mismatch in the time scale of the incorporated features, and the features present more complex change patterns and trends. Through time series feature extraction, patterns and regularities in the data, such as trends, periodicity and noise, can be found. The accuracy and reliability of the algorithm can be significantly improved by applying the extracted information features as independent variables in the model [24]. In this study, 21 statistical indicators such as the mean, variance, and waveform factor of each underlying variable are mainly extracted.

After feature extraction, the number of features will be significantly increased, if directly input into the prediction model there will be data imbalance phenomenon, thus increasing the risk of overfitting. Therefore, feature

Table 2 Basic features and indicators for extracting information

Category	Specific Index
JC_Baidu Idex	BCG Vaccine、 Fever、Cough、Tuberculosis、Isoniazid
JC_Pollutant	$PM_{2.5}$、PM_{10}、SO_2、CO、NO_2、O_3
JC_Climate	T、Po、P、Pa、U、Tn、Tx、VV、Td
JC_Stock	Anhui Plate、Vaccine Plate (Opening、Highest、Lowest、Closing、Increase、Amplitude)
Statistical Information	maximum、minimum、mean、peak-peak、rectification mean、effective value、peak、variance、standard deviation、kurtosis, skewness、root mean square、waveform factor、peak factor、pulse factor、margin factor、center of gravity frequency、mean square frequency、frequency variance、band energy、relative power spectrum entropy

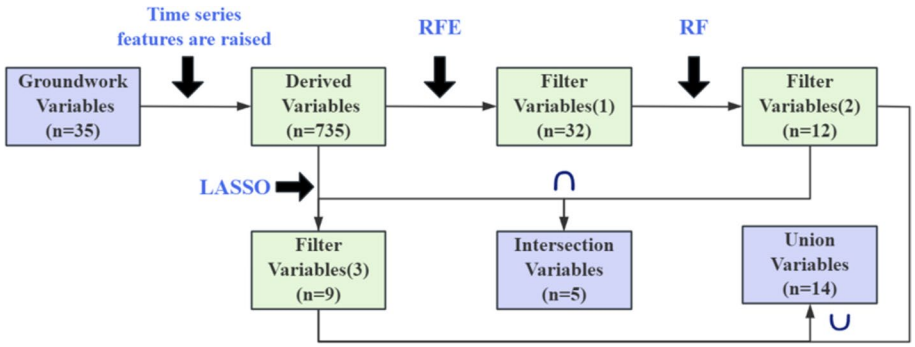

Fig. 2 Flowchart of feature extraction and filtering

screening is an essential part of the process, and commonly used feature screening methods include: Recursive Feature Elimination (RFE), Random Forest and Lasso regression (see Fig. 2).

Recursive Feature Elimination (RFE) is mainly used to select the optimal subset of features by recursively training the model and eliminating the features. The method trains the model with the current optimal feature subset during each iteration and evaluates the model performance on new data. Then, based on the model performance and the importance of the features, the least important features are selected and removed from the feature set. This process is repeated recursively until a predefined stopping condition is met, such as reaching the maximum number of iterations or reaching the desired number of features [25].

Random forest is an integrated learning method that improves prediction performance by constructing multiple decision trees and combining their predictions. In random forest, feature selection is an important step that helps us to find the features that have the greatest impact on the prediction results. For each decision tree, during the construction process, Random Forest will use the self-help method (bootstrap) to extract samples from the original dataset, and then perform feature selection for each sample; after the training of all the decision trees is completed, Random Forest will use the voting method (voting) to synthesize the prediction results of each decision tree [26].

Lasso regression is a classical feature selection and regression analysis method, which achieves feature selection by adding an L1 regular term to the loss function so that some coefficients are close to zero. Similarly, for feature selection, Lasso regression is able to deal with the problem of multicollinearity, and Lasso regression is interpretable and can output the importance of each feature. It can also fulfill its unique advantages when dealing with large-scale datasets and high-dimensional data [27, 28].

Model building and optimization

There are more method models for forecasting time series data, the common ones are STL model, ARIMA model, and Holt-Winters model, etc. [29] With the rise of machine learning, more and more machine learning models are also used to forecast and analyze the time series data, and better results have been achieved [30].

SARIMA (Seasonal Autoregressive Integrated Moving Average) model is a classical model used for time series forecasting, which is composed of seasonal

autoregressive model (SAR), differential integration (I), and moving average (MA).The SARIMA model is mainly used to forecast the time series by capturing the seasonality and trend to predict future values.The parameters of SARIMA model can be determined by autocorrelation function (ACF) and partial autocorrelation function (PACF) [31]. Its basic form can be expressed as:

$$\text{SARIMA } (p\ d\ q)(P\ D\ Q)_s$$

Model is a time series prediction model based on neural network, which was proposed by Japanese scholar Nishiguchi Taro in 2017. The neural network model combines neural network and autoregressive model (AR). The neural network is used to learn the latent structure of time series, and the AR model is used to control the long-term dependence of time series. The main idea of the model is to decompose the time series data into three parts: trend, seasonality and noise, and then use the neural network to learn the pattern of trend and seasonality, and use the autoregressive model to model the noise component [32]. Based on the neural network model, for multivariate prediction, the researchers constructed a neural network model, Combining the external input with the internal state to predict the output at the next time point, the neural network model model is usually used to solve the system problem with nonlinear dynamic characteristics, has good generalization ability, and can handle complex time series data. [33]. Long short-term memory) neural network is a special neural network (recurrent neural network, recurrent neural network) structure, First proposed by Hawklett and Schmidhuber in 1997, long short-term memory has better long-term dependence characteristics than traditional neural networks, and can effectively deal with long-term dependence in time series data (see Fig. 3). The main features of the least square method include: 1. Memory cell: Memory cell contains a special structure called "memory cell," which is used to store and update the information about the sequence; 2. Gated structure: The system contains three gates (input gate, forget gate and output gate), which are used to control the flow of information and memory update; 3. Vanishing gradient problem: By introducing memory unit and gating structure in the least square model, the vanishing gradient problem in the network is effectively solved, thus improving the training effect of the model [34].

Particle Swarm Optimization (PSO) is an optimization algorithm based on the behavior of a flock, first proposed by Eberhart and Kennedy in 1995. The algorithm is inspired by the foraging behavior of bird flocks and finds the optimal solution by simulating the interaction of individuals in the flock. The particle swarm optimization algorithm considers both global and individual optimal solutions. Each particle uses the information of both the global optimal solution and the individual optimal solution during the updating process, which enables it to converge to the optimal solution faster. In addition, exploration and convergence are balanced by inertia weights and learning factors, thus maintaining the diversity of the population during the search process. This helps to avoid the algorithm from falling into local optimal solutions. Particle swarm optimization algorithms have been widely used in a variety of optimization problems, such as function optimization, constrained optimization, combinatorial optimization, and machine learning [35].

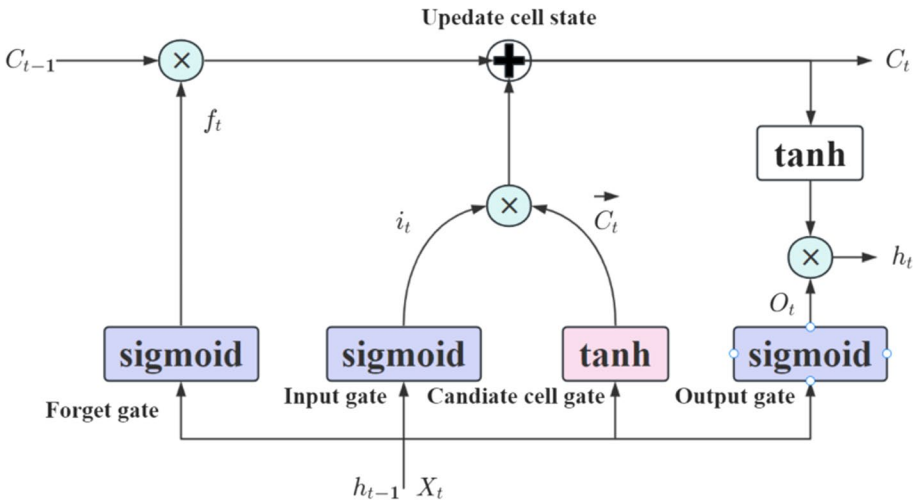

Fig. 3 Schematic diagram of the LSTM model

Results

In the results section, the analysis is first carried out to evaluate the model performance by univariate autoregressive modeling. The model evaluation indexes mainly include MSE (Mean Square Error), MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), and R^2 (Coefficient of Determination), and the smaller the first four items are, the better the model performance is, and the closer the R^2 is to 1, the better the model prediction performance is.The evaluation of the results is based on the size of the R^2 value of the test set as the main basis, and the rest of the test set indicators as the secondary basis.

Univariate time series modeling

After collecting data on the number of TB cases from December 2013 to October 2023 in Anhui Province, the serial data were modeled using the SARIMA model as well as the NAR model, respectively. The model was constructed using the data from December 2013 to October 2022 to predict the number of monthly incidence from November 2022 to October 2023, and the results were compared with the real results, which are shown in Fig. 4 and Table 3.

Among them, according to the autocorrelation and partial autocorrelation regression results, using the AIC value to compare with the BIC value, the final will of the SARIMA model has the p, d, and q of 1, 0, and 1, respectively, and the values of P, D, and Q of 1, 1, and 1, respectively. the NAR model has the step size of 2, the number of neurons of 20, and the trainFcn is set to Levenberg–Marquardt backpropagation.

From the above, it can be seen that both the SARIMA model and the NAR model have poor predictive performance, and the inclusion of multivariate for infectious disease analysis is an important part to improve the accuracy of the model.

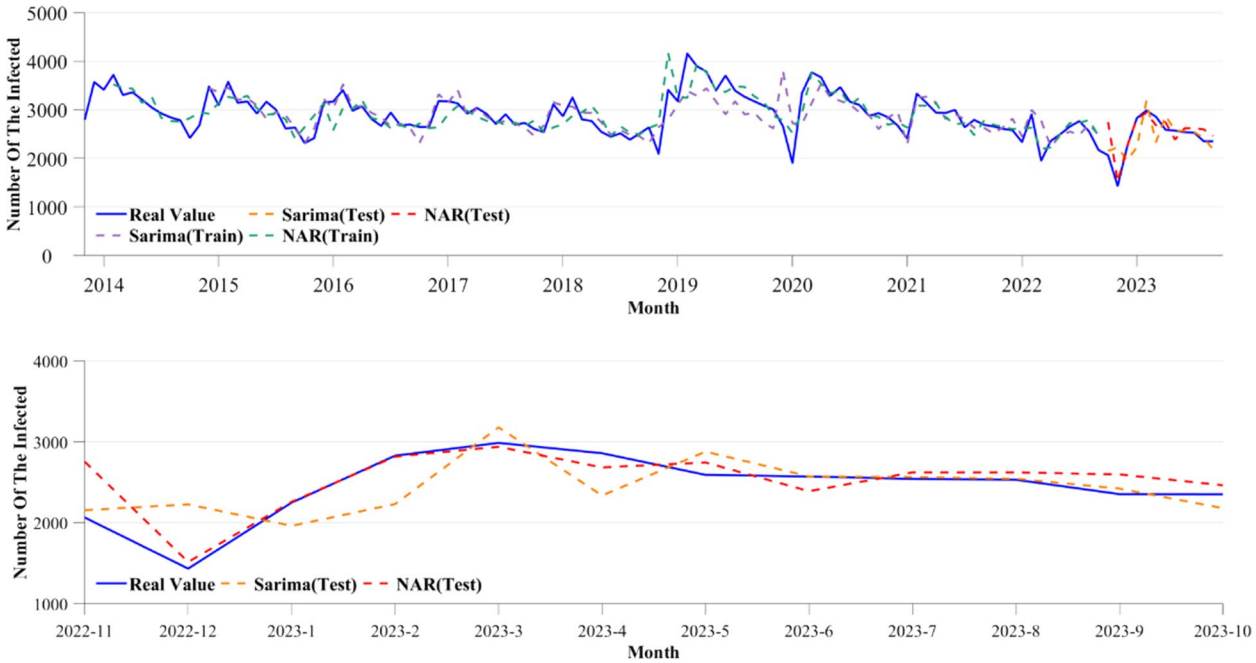

Fig. 4 Model fitting results (SARIMA, NAR)

Table 3 Model evaluation indexes

Model	Data Set	MSE	MAE	RMSE	MARE	R^2
SARIMA	Training set	30.7440	206.6809	298.0741	7.2432	0.4924
	Test set	102.2350	253.8333	354.1525	11.7597	0.2041
NAR	Training set	24.4480	178.9808	249.3214	6.2521	0.6381
	Test set	67.6513	156.4167	234.3511	6.8784	0.6515

The training set model evaluation results incorporate the predictions from the validation data

Evaluation of feature screening methods

As shown in Table 2, 21 monthly statistical features can be extracted for each base variable to achieve consistency between the independent and dependent variables on the time scale. After incorporating 35 basic variables for feature extraction derived as 735 features, 12 important variables are screened out through RFE-RF method, and the data set named S1 is constructed with the number of morbidity; 9 important variables are screened out by using LASSO, and the data set named S2 is constructed; the intersection variables of the two screening processes are 5, and the S3 data set is constructed with the dependent variable; and the concatenated set of variables, totaling 14, is formed. S4 dataset is formed. The NARX model is used to predict and analyze these four data sets and compare the final results. The division of the data sets is consistent with 2.1, and the model parameters are adjusted by using five-fold cross-validation. In order to facilitate the adjustment of model parameters, the data from November 2021 to October 2022 were set as the validation set data, and the model parameters were adjusted using the validation set model evaluation indicators. The results are shown in Fig. 5 and Table 4.

The step size of the NARX model is 12, the number of Shenjing elements is 50, and trainFcn is also set to Levenberg–Marquardt backpropagation. The results of correlation analysis show that the S4 dataset has the best effect, and the corresponding variables are 14. The results of correlation analysis show that the correlation between variables is small and there is no collinearity (see Fig. 6).

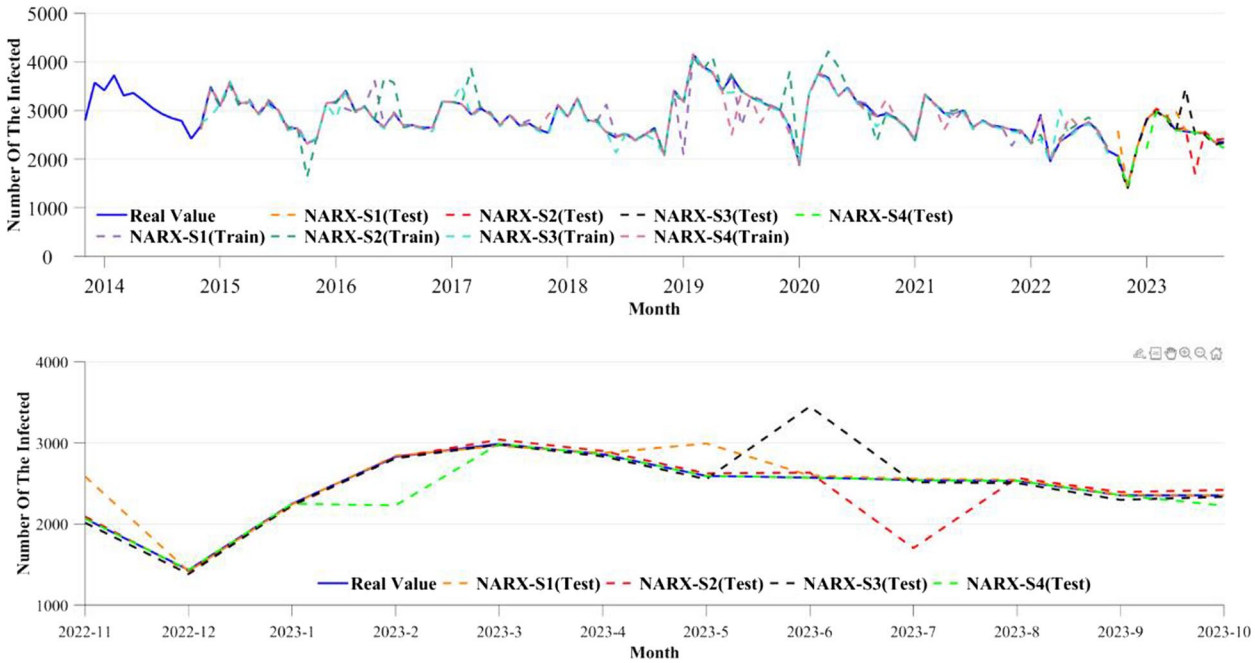

Fig. 5 Model fitting results of each dataset

Table 4 Model evaluation indexes

Data Type	Data Set	MSE	MAE	RMSE	MARE	R^2
S1	Training set	18.1393	61.8000	176.8000	2.0872	0.8201
	Test set	54.9376	88.3333	190.3094	3.8754	0.7702
S2	Training set	24.2809	93.6737	236.6605	3.2776	0.6777
	Test set	70.8886	103.5833	245.5652	4.1237	0.6173
S3	Training set	13.6990	59.6105	133.5217	2.0894	0.8974
	Test set	73.6139	101.0833	255.0060	4.1138	0.5873
S4	Training set	14.3776	39.1895	140.1352	1.4123	0.8870
	Test set	**56.7019**	**62.2500**	**196.4211**	**2.2124**	**0.7552**

The training set model evaluation results incorporate the predictions from the validation data

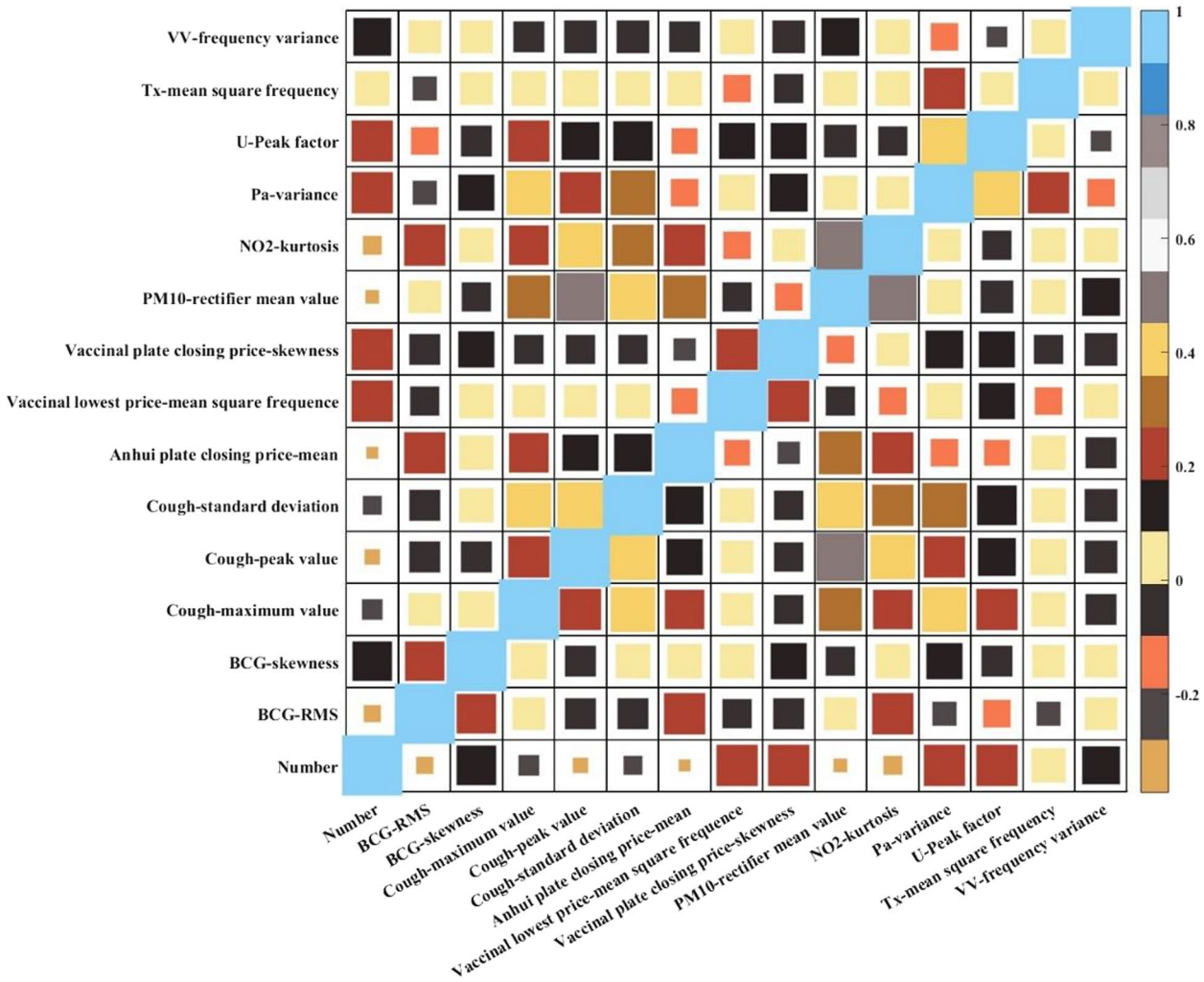

Fig. 6 Results of correlation analysis

Number indicates the number of TB cases, and the rest of the features are expressed using A-B, where A indicates the specific indicator and B indicates the extracted information, e.g. BCG-RMS indicates the valid values extracted from the BCG monthly index data

Multi-input time series modeling

Using the S4 dataset constructed in Part 2.2 for modeling analysis, PSO-NARX, MiLSTM and PSO-MiLSTM were constructed respectively. The training set, validation set and test set are divided in the same way as in Sect. "Feature extraction and screening", and the results of the validation set are used as the objective function of the optimisation algorithm. The result of five-fold cross-validation was used as the objective function. The results are shown in Fig. 7 and Table 5.

After combining the intelligent algorithm for hyperparameter optimisation, the complexity of MiLSTM model parameter adjustment can be effectively solved through cross-validation, and PSO-MiLSTM achieves the most excellent prediction performance, with step size by 12, maximum number of iterations by 236; initial learning rate by 0.001; learning rate decline factor by 0.01. The model test set MSE is 42.3555, MAE is 59.3333, RMSE is 146.7237, MAPE is 2.1133, and R^2 is 0.8634.

Discussion

The spread and development of infectious diseases can lead to serious health problems and death. For example, the new coronavirus (COVID-19) can cause serious diseases such as pneumonia, acute respiratory distress syndrome, multi-organ failure, and even death. It can also have an impact on many aspects such as socio-economics, public safety, education, and tourism [36]. Mycobacterium tuberculosis is highly contagious and easily spreads among the population, especially in confined and crowded environments, such as schools,

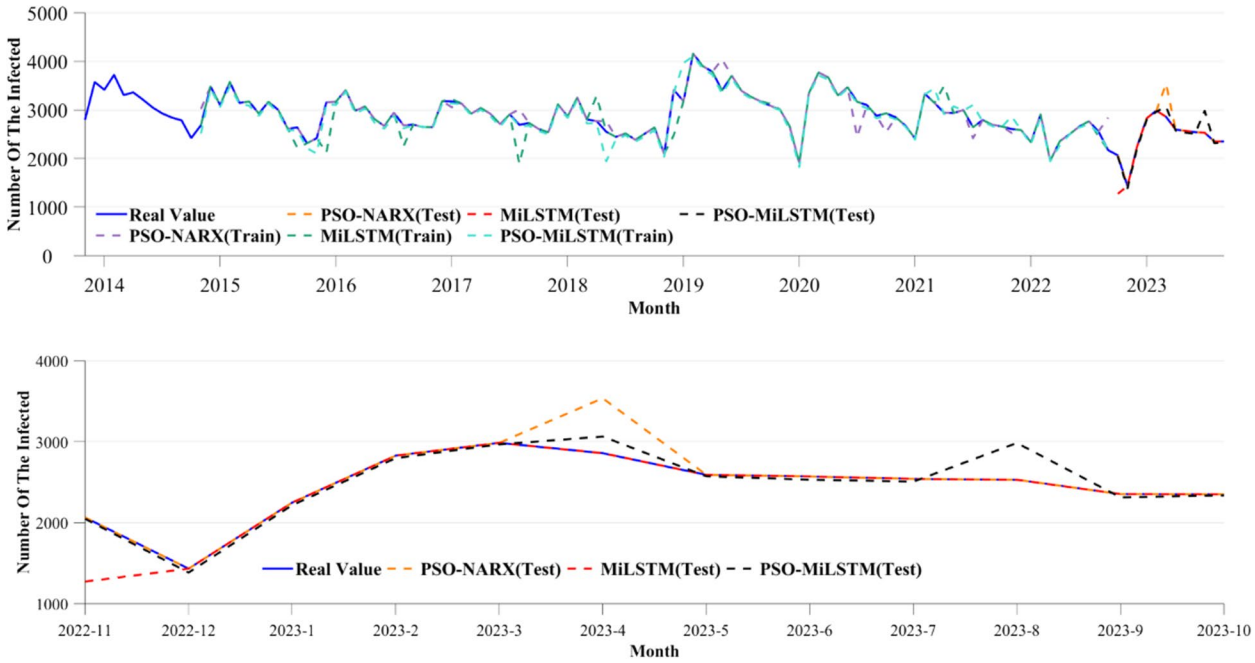

Fig. 7 Fitting results of each model

Table 5 Evaluation indexes of each model

Model	Data Set	MSE	MAE	RMSE	MARE	R^2
PSO-NARX	Training set	15.5068	35.6211	151.1420	1.1616	0.8685
	Test set	50.9920	60.3333	176.6414	2.2082	0.8020
MiLSTM	Training set	19.7280	51.1158	192.2843	1.7405	0.7872
	Test set	56.0133	66.0200	228.6307	3.1977	0.6683
PSO-MiLSTM	Training set	13.7671	71.4632	134.1847	2.5264	0.8964
	Test set	**42.3555**	**59.3333**	**146.7237**	**2.1133**	**0.8634**

The training set model evaluation results incorporate the predictions from the validation data

factories, and prisons. The prevention and treatment of TB consumes a large amount of medical resources and funds, and imposes a heavy economic burden on individuals, families, and society. TB epidemics may lead to labor shortages and decreased production capacity, thus affecting socioeconomic stability and development [37]. Forecasting and analyzing the development trend of tuberculosis epidemics in various geographic regions can provide an adequate objective basis for the development of preventive and control measures. In the analysis process, it is crucial to establish a stable and accurate prediction model by utilizing various methods [38].

With the increase of external interfering factors, using only historical data of infectious diseases to predict future trends has greater limitations. Data from multiple sources were used for feature analysis, and the variables with the highest correlation with the number of incidence cases were screened as input data, which were combined with the autoregressive part of the time-series data to fit the number of TB incidence cases through the MiLSTM model, which is conducive to improving the accuracy of the model. Among them, the steps of feature extraction, feature screening, base model and parameter optimization are the key links to improve the accuracy and generalization of the model, and the diversification of feature extraction can achieve the purpose of making full use of the data, but it will bring the problem of too high feature dimension, which needs to be solved by choosing the appropriate feature screening method. The complexity of the deep learning model structure can more accurately capture the trend of the dependent variable, but compared with the ordinary model, the number of hyperparameters increases significantly, and it is necessary to choose the appropriate optimization algorithm for automatic optimization [38–40].

Therefore, in the analysis and prediction of tuberculosis incidence in Anhui Province, we conducted a comparative study to evaluate two univariate input models, four feature screening approaches and three multi-input models. On this basis, we proposed the RRL feature screening strategy. To address the difficulty in tuning the model performance due to the increase in parameters, the MiLSTM model performance was further improved using the PSO optimisation algorithm. Finally, we successfully constructed the optimal model PSO-MiLSTM and achieved the best prediction performance on the test set.

Although this study is reasonable and rigorous in both design and execution, there is still some room for improvement and enhancement. We look forward to future research directions. First, a multi-dimensional and broad field analysis should be conducted in terms of features such as spatial distribution, population distribution and causative factors for better implementation of early warning. Second, theoretical research should be conducted on the principles of formulaic methods such as MSE, huber loss function, and quantile loss function in terms of loss function construction, so as to construct a loss function that conforms to the infectious disease data, and to further improve the prediction effect of the model. Finally, the prediction effect of the model should be verified on multiple infectious disease datasets to verify its generalization ability.

In summary, this study uses data on atmospheric pollutants and climate as external input variables, combined with the time-series characteristics of the number of infectious disease cases, to carry out trend prediction analysis of the number of tuberculosis infections in Anhui Province. Based on the NAR model and NARX model, the S4 feature dataset consistent with the predictive analysis of infectious diseases was constructed by performing feature extraction and feature screening, and the excellence of the predictive performance of PSO-MiLSTM was revealed by empirical studies. Therefore, the feature engineering method as well as the prediction model proposed in this study can effectively analyse the development trend of the occurrence of tuberculosis epidemics, and by monitoring the change trend of the screened features, it can accurately determine the change trend of the number of tuberculosis cases in the next 1 month, and provide a powerful guide for the prevention and control of other infectious diseases.

In future studies, it is planned to incorporate data from more geographical areas, including regions with different climatic, economic and socio-cultural backgrounds, to test the stability and accuracy of the model in different environments. It is also planned to apply the model to more types of infectious diseases, such as other respiratory diseases, gastrointestinal diseases, and insect-borne infectious diseases, in order to explore the potential value and application of the model in the prediction of different diseases.

Conclusion

In summary, this study utilized data on atmospheric pollutants and climate as external input variables, and combined the time-series characteristics of the number of infectious disease incidence with the trend prediction analysis of the number of tuberculosis infections in Anhui Province. On the basis of NAR model and NARX model, the S4 feature dataset conforming to the predictive analysis of infectious diseases was constructed by performing feature extraction and feature screening, and the excellence of the predictive performance of PSO-MiLSTM was revealed by empirical studies. Therefore, the feature engineering method as well as the prediction model proposed in this study can effectively analyze the development trend of the occurrence of tuberculosis epidemics, and by monitoring the change trend of the screened features, it can accurately determine the change trend of the number of tuberculosis cases in the next 1 month, and provide a powerful guide for the prevention and control of other infectious diseases.

Authors' contributions
All authors contributed to the study conception and design. Methodology and Formal analysis were performed by Yan Zhang, Huan Ma and Qing Xia. Material preparation and data collection were performed by Shasha Wu, Jing Meng, Panpan Zhu and Zhilong Guo. The first draft of the manuscript was written by Yan Zhang, Huan Ma. Project administration, Supervision and Writing—review & editing were performed by Jing Hou and Hua Wang. All authors read and approved the final manuscript.

Funding
This study was supported by Scientific Research Project of Anhui Provincial Health Commission under AHWJ2022b040.

Data availability
The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 January 2024 Accepted: 19 September 2024
Published online: 26 October 2024

References

1. World Health Organization. Tuberculosis. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/tuberculosis. Accessed Nov 2023.
2. Meyers LA. Contact network epidemiology: bond percolation applied to infectious disease prediction and control. Bull Am Math Soc. 2007;44:63–86.
3. Dimitrov NB, Meyerss LA. Mathematical approaches to infectious disease prediction and control. INFORMS Tutor Oper Res. 2010;7:1–25.
4. Hyndman RJ, Khandakar D. Automatic time series forecasting using regression models. J Mach Learn Res. 2008;9(3):253–85.
5. Kim L, Fast SM, Markuzon N. Incorporating media data into a model of infectious disease transmission. PLoS ONE. 2019;14(2):e0197646. https://doi.org/10.1371/journal.pone.0197646.
6. Amendolara AB, Sant D, Rotstein HG, Fortune E. LSTM-based recurrent neural network provides effective short term flu forecasting. BMC Public Health. 2023;23(1):1788. https://doi.org/10.1186/s12889-023-16720-6.
7. Moursi ASA, El-Fishawy N, Djahel S, Shouman MA. Enhancing PM2.5 Prediction Using NARX-Based Combined CNN and LSTM Hybrid Model. Sensors (Basel). 2022;22(12):4418. https://doi.org/10.3390/s22124418.
8. Kim J, Ahn I. Infectious disease outbreak prediction using media articles with machine learning models. Sci Rep. 2021;11(1):4413. https://doi.org/10.1038/s41598-021-83926-2.
9. Liu H, Manzoor A, Wang C, Zhang L, Manzoor Z. The COVID-19 outbreak and affected countries stock markets response. Int J Environ Res Public Health. 2020;17(8):2800.
10. Arthi V, Parman J. Disease, downturns, and wellbeing: Economic history and the long-run impacts of COVID-19. Explor Econ Hist. 2021;79:101381. https://doi.org/10.1016/j.eeh.2020.101381.
11. Liu YY, Viboud C. Climate change and infectious diseases: what can we expect? Lancet Infect Dis. 2018;18(12):1251–2.
12. Wu Y, Huang C. Climate change and vector-borne diseases in china: a review of evidence and implications for risk management. Biology (Basel). 2022;11(3):370. https://doi.org/10.3390/biology11030370.
13. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, Del Valle S, Manore CA. Google Health Trends performance reflecting dengue incidence for the Brazilian states. BMC Infect Dis. 2020;20(1):252. https://doi.org/10.1186/s12879-020-04957-0.
14. Rui Zhang, Chengcheng Gao, Xicheng Chen, Fang Li, Dong Yi, Yazhou Wu. Genetic algorithm optimised Hadamard product method for inconsistency judgement matrix adjustment in AHP and automatic analysis system development, Expert Systems with Applications, Volume 211, 2023, 118689, ISSN 0957–4174, https://doi.org/10.1016/j.eswa.2022.118689.
15. Arora T, Dhir R. Correlation-based feature selection and classification via regression of segmented chromosomes using geometric features. Med Biol Eng Comput. 2017;55(5):733–45. https://doi.org/10.1007/s11517-016-1553-2.
16. Qiu S, Cui X, Ping Z, Shan N, Li Z, Bao X, Xu X. Deep learning techniques in intelligent fault diagnosis and prognosis for industrial systems: a review. Sensors (Basel). 2023;23(3):1305. https://doi.org/10.3390/s23031305.
17. Pintas JT, Fernandes LAF, Garcia ACB. Feature selection methods for text classification: a systematic literature review. Artif Intell Rev. 2021;54:6149–200. https://doi.org/10.1007/s10462-021-09970-6.
18. Ein Shoka AA, Alkinani MH, El-Sherbeny AS, El-Sayed A, Dessouky MM. Automated seizure diagnosis system based on feature extraction and channel selection using EEG signals. Brain Inform. 2021;8(1):1. https://doi.org/10.1186/s40708-021-00123-7.
19. Karthik KV, Rajalingam A, Shivashankar M, Ganjiwale A. Recursive feature elimination-based biomarker identification for open neural tube defects. Curr Genomics. 2022;23(3):195–206. https://doi.org/10.2174/1389202923666220511162038.
20. Deviaene M, Testelmans D, Borzee P, Buyse B, Huffel SV, Varon C. Feature selection algorithm based on random forest applied to sleep apnea detection. Annu Int Conf IEEE Eng Med Biol Soc. 2019;2019:2580–3. https://doi.org/10.1109/EMBC.2019.8856582.
21. R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 2016, pp. 18–20, https://doi.org/10.1109/ICACA.2016.7887916.
22. Wang H, Tian CW, Wang WM, Luo XM. Time-series analysis of tuberculosis from 2005 to 2017 in China. Epidemiol Infect. 2018;146(8):935–9. https://doi.org/10.1017/S0950268818001115.
23. Kazijevs M, Samad MD. Deep imputation of missing values in time series health data: A review with benchmarking. J Biomed Inform. 2023;144:104440. https://doi.org/10.1016/j.jbi.2023.104440.
24. Herff C, Extracting KDJ, Features from Time Series. 22. In: Kubben P, Dumontier M, Dekker A, editors. Fundamentals of clinical data science. Cham (CH): Springer; 2018. p. 2019.
25. Bradley A, Fayyad U. Regression and time series model selection: A combined approach. J Forecast. 1998;17(4):337–54.
26. Zhang H, Zhu J. Random forests based on variable importance: Algorithm and applications to feature selection. Expert Syst Appl. 2009;36(4):7200–11.
27. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net,Journal of the Royal Statistical Society Series B. Stat Methodol. 2005;67(2):301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.
28. Hothorn T, Bühlmann P, Gneiting T. Variable selection for general regression models. J Stat Plan Inference. 2006;136(4):1194–206.
29. Krymova E, Béjar B, Thanou D, Sun T, Manetti E, Lee G, et al. Trend estimation and short-term forecasting of COVID-19 cases and deaths worldwide. Proc Natl Acad Sci U S A. 2022;119(32):e2112656119. https://doi.org/10.1073/pnas.2112656119.
30. Effrosynidis D, Spiliotis E, Sylaios G, Arampatzis A. Time series and regression methods for univariate environmental forecasting: An empirical evaluation. Sci Total Environ. 2023;875:162580. https://doi.org/10.1016/j.scitotenv.2023.162580.
31. Duangchaemkarn K, Boonchieng W, Wiwatanadate P, Chouvatut V. SARIMA Model Forecasting Performance of the COVID-19 Daily Statistics in Thailand during the Omicron Variant Epidemic. Healthcare (Basel). 2022;10(7):1310. https://doi.org/10.3390/healthcare10071310.
32. Zhipeng Shen, Yuanming Zhang, Jiawei Lu, Jun Xu, Gang Xiao, A novel time series forecasting model with deep learning, Neurocomputing, Volume 396, 2020, Pages 302–313, ISSN 0925–2312, https://doi.org/10.1016/j.neucom.2018.12.084.
33. Di Nunno F, Race M, Granata F. A nonlinear autoregressive exogenous (NARX) model to predict nitrate concentration in rivers. Environ Sci Pollut Res Int. 2022;29(27):40623–42. https://doi.org/10.1007/s11356-021-18221-8.
34. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 2019;31(7):1235–70. https://doi.org/10.1162/neco_a_01199.
35. Li P, Yang J. PSO Algorithm-based design of intelligent education personalization system. Comput Intell Neurosci. 2022;2022:9617048. https://doi.org/10.1155/2022/9617048. Retraction.In:ComputIntellNeurosci.2023Aug,23(2023),pp.9780681.
36. Sharma A, Ahmad Farouk I, Lal SK. COVID-19: A review on the novel coronavirus disease evolution, transmission, detection, control and prevention. Viruses. 2021;13(2):202. https://doi.org/10.3390/v13020202.
37. Chakaya J, Petersen E, Nantanda R, Mungai BN, Migliori GB, Amanullah F, Lungu P, Ntoumi F, Kumarasamy N, Maeurer M, Zumla A. The WHO Global Tuberculosis 2021 Report - not so good news and turning the tide back to End TB. Int J Infect Dis. 2022;124(Suppl 1):S26–9. https://doi.org/10.1016/j.ijid.2022.03.011.
38. Hie BL, Yang KK. Adaptive machine learning for protein engineering. Curr Opin Struct Biol. 2022;72:145–52. https://doi.org/10.1016/j.sbi.2021.11.002.
39. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019;28(2):73–81. https://doi.org/10.1080/13645706.2019.1575882.
40. Hassabis D, Kumaran D, Summerfield C, Botvinick M. Neuroscience-inspired artificial intelligence. Neuron. 2017;95(2):245–58. https://doi.org/10.1016/j.neuron.2017.06.011.

Publisher's Note