BMJ Open

Predicting the impact of climate change on the re-emergence of malaria cases in China using LSTMSeq2Seq deep learning model: a modelling and prediction analysis study

Eric Kamana , Jijun Zhao, Di Bai

To cite: Kamana E, Zhao J, Bai D. Predicting the impact of climate change on the re-emergence of malaria cases in China using LSTMSeq2Seq deep learning model: a modelling and prediction analysis study. *BMJ Open* 2022;**12**:e053922. doi:10.1136/bmjopen-2021-053922

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-053922).

Received 27 May 2021
Accepted 25 February 2022

Complexity Science Institute, School of Automation, Qingdao University, Qingdao, China

Correspondence to
Jijun Zhao; jjzhao@qdu.edu.cn

ABSTRACT

Objectives Malaria is a vector-borne disease that remains a serious public health problem due to its climatic sensitivity. Accurate prediction of malaria re-emergence is very important in taking corresponding effective measures. This study aims to investigate the impact of climatic factors on the re-emergence of malaria in mainland China.

Design A modelling study.

Setting and participants Monthly malaria cases for four *Plasmodium* species (*P. falciparum*, *P. malariae*, *P. vivax* and other *Plasmodium*) and monthly climate data were collected for 31 provinces; malaria cases from 2004 to 2016 were obtained from the Chinese centre for disease control and prevention and climate parameters from China meteorological data service centre. We conducted analyses at the aggregate level, and there was no involvement of confidential information.

Primary and secondary outcome measures The long short-term memory sequence-to-sequence (LSTMSeq2Seq) deep neural network model was used to predict the re-emergence of malaria cases from 2004 to 2016, based on the influence of climatic factors. We trained and tested the extreme gradient boosting (XGBoost), gated recurrent unit, LSTM, LSTMSeq2Seq models using monthly malaria cases and corresponding meteorological data in 31 provinces of China. Then we compared the predictive performance of models using root mean squared error (RMSE) and mean absolute error evaluation measures.

Results The proposed LSTMSeq2Seq model reduced the mean RMSE of the predictions by 19.05% to 33.93%, 18.4% to 33.59%, 17.6% to 26.67% and 13.28% to 21.34%, for *P. falciparum, P. vivax, P. malariae*, and other *plasmodia*, respectively, as compared with other candidate models. The LSTMSeq2Seq model achieved an average prediction accuracy of 87.3%.

Conclusions The LSTMSeq2Seq model significantly improved the prediction of malaria re-emergence based on the influence of climatic factors. Therefore, the LSTMSeq2Seq model can be effectively applied in the malaria re-emergence prediction.

Strengths and limitations of this study

- The use of climatic factors has proven to be effective predictors for malaria incidence and significantly affect the proposed long short-term memory sequence-to-sequence (LSTMSeq2Seq) model in capturing seasonal patterns and trends and predicting malaria incidence.
- It is hard for a typical machine learning model to predict long-term dependencies, and it is even difficult for a single LSTM to capture key past events and use them to predict future values. By combining specialised LSTM cells that can forecast multiple time steps rather than having one multitasking cell, LSTMSeq2Seq solved this problem.
- The LSTMSeq2Seq takes more time for training than other employed deep learning models. To train the LSTMSeq2Seq from scratch for all 31 provinces takes 2 weeks for four types of *Plasmodium* used in our study. Whereas other models take a few hours to days to train them using malaria cases and data of meteorological variables. In many provinces, LSTM was seven times faster than the LSTMSeq2Seq model. However, the impact is not significant in provinces with fewer malaria cases.
- We could not obtain accurate predictions in some provinces by using any model in this study, due to the lack of other relevant potential non-climatic factors.

INTRODUCTION

Malaria is a vector-borne infectious disease caused by the parasitic protozoans of the genus *Plasmodium* such as *Plasmodium falciparum* (*P. falciparum*), *Plasmodium ovale* (*P. ovale*), *Plasmodium vivax* (*P. vivax*), *Plasmodium simium* (*P. simium*), *Plasmodium knowlesi* (*P. knowlesi*), and *Plasmodium cynomolgi* (*P. cynomolgi*). Governments, health organisations and scientific research institutions all over the world have made significant efforts on malaria control measures and elimination

programmes. Despite the huge progress in reducing malaria cases and deaths, malaria remains life-threatening to global health mainly in Africa, Asia and America continents due to its sensitivity to environmental and climatic changes.

According to the World Malaria Report 2020 published by WHO, a total of 229 million malaria cases and 409 000 deaths were reported worldwide in 2019.[1] Most of the malaria cases (93%) and malaria deaths (94%) occurred in the WHO African region, while the other WHO regions shared the remaining percentages.[1] Despite remarkable progress, the global gains in fighting malaria disease have levelled off in recent years, and many high burdens have been losing ground. The combat against malaria had reached a crossroad.[1] The world did not meet the milestones that could lower malaria cases and mortality by 90% by 2030. Without a massive coordinated action, the world is unlikely to meet the WHO's Global Technical Strategy for malaria 2016–2030 targets.[2] The COVID-19 pandemic has complicated the malaria picture even further, according to the WHO modelling analysis. The recent WHO report features a particular section on the COVID-19 pandemic and malaria, which could potentially double the number of malaria deaths in the WHO African region due to the disruptions to insecticide-treated net campaigns and the interruptions to access to antimalarial medicines.

Historically, malaria was one of the most prevalent parasitic diseases in the People's Republic of China. However, through many years of combatting malaria, the Chinese government achieved remarkable progress in reducing malaria incidences through effective treatment and vector control measures. Vector control measures include reducing mosquito breeding grounds, implementing antimalaria grassroots campaigns.[3] In 2010, the Chinese government launched the National Malaria Elimination Program.[4–6] Indigenous malaria cases dramatically decreased to zero in 2017, which marked China among 21 countries with the potential of achieving a malaria eradication plan certified by WHO.[7] However, imported *P. falciparum* malaria cases increased in many provinces, which poses a challenge to achieve malaria-free status and might cause another situation of malaria re-emergence that has been identified in some countries.[8 9] A surveillance system in China is used to detect imported malaria cases but may miss some. Mosquitos are still out there with the ability to transmit the undetected imported malaria cases.

The re-emergence of malaria happened in Anhui and Henan provinces at the beginning of the 21st century. The re-emergence was due to climatic change, population movement, *Anopheles* abundance increase as well as mosquito's drug resistance.[10 11] Malaria outbreaks and re-emergence in the Huang-Huai River region happened due to the increase of *Anopheles sinensis* (*An. sinensis*). There was a high relationship between the re-emergence of *P. vivax* and an increase in the vectorial capacity of *An. sinensis*.[12 13] Climatic conditions as the concerning factors in this study have contributed to the re-emergence of malaria by providing favourable conditions for the breeding and survival of mosquitoes.[14] Numerous studies attempted to identify and assess the impact of climatic factors on malaria incidence in China.[15–17] Some studies reported that the intra-annual variation in malaria cases might associate with changes in ambient temperature, precipitation, relative humidity, wind direction, sunshine duration and wind speed. Nevertheless, the findings were inconsistent in key factors observed and the corresponding effects estimated. Zinszer *et al*[18] reviewed previously published studies related to the different approaches and factors used to predict malaria incidence. Most of the predictors were related to climate factors. Statistical, mathematical, machine learning and deep learning models have been applied to these climate predictors to improve the forecast accuracy of malaria incidence. Wang *et al*[19] proposed an ensemble approach of traditional time series and deep learning models to improve the prediction performance of malaria incidence using malaria and climate data in Yunnan province. The study applied time series and deep learning models such as autoregressive integrated moving average (ARIMA), seasonal and trend decomposition using loess—integrated moving average (STL+ARIMA), backpropagation artificial neural network and long short-term memory (LSTM) network separately on the prepared data. Different evaluation methods were used to compare the prediction accuracy of methods. Gradient-boosting regression trees combine different models and are trained using climate data and malaria incidence. The model outperforms traditional time series and deep learning methods. Nkiruka *et al*[20] proposed a machine learning system to assess the association between climatic factors and malaria incidence and found that rainfall, surface radiation and temperature affect the outbreak of malaria disease.

The relationship between malaria incidence and climatic factors is complex and cannot easily fit the classical forecasting approaches and machine learning algorithms. To reduce the complexity of this relationship by predicting malaria incidence with remarkable performance, deep learning models offer more advantages in the healthcare field by interacting with training data. Deep learning models give more accurate predictions compared with the statistical and mathematical approaches. Through deeper hidden layers, deep learning methods help us to gain unprecedented insights into care processes, diagnostics and forecasting and can make meaning from medical data. Deep learning models were applied to the prediction of directly transmitted infectious diseases.[21] Some of the advanced deep learning models like LSTM and gated recurrent unit (GRU) recurrent neural networks with a large number of discrete time steps have been used in predicting infectious diseases like influenza, dengue incidence and hand, foot and mouth disease. LSTM model outperformed other machine learning models by achieving accuracy prediction and lower root mean squared error (RMSE).[22–26] In this research, we

identified and assessed climatic factors as predictors that may contribute to the re-emergence of malaria disease in China. We used climate factors with malaria incidence to train our constructed deep learning sequence-to-sequence model (LSTMSeq2Seq) and then evaluated its performance by predicting the re-emergence of malaria disease in China.

METHODOLOGY

Patient and public involvement

No patient involved.

Data collection and data preprocessing

We collected monthly malaria cases in all 31 provinces in China from January 2004 to December 2016. The data set contains four classes of *Plasmodium* species that is *P. falciparum*, *P. vivax*, *P. malariae* and other *Plasmodium* species. The *plasmodium* species category named other could be *P.ovale*, *P. knowlesi* or unidentified species type. Malaria cases for all 31 provinces of mainland China were obtained from the Chinese Center for Disease Control and Prevention (www.phsciencedata.cn)[27] which provides the database for infectious diseases. The meteorological data of these 31 provinces were obtained from the China Meteorological data service centre (http://data.cma.cn/en).[28] A total of 10 meteorological variables (ie, pressure, average temperature, maximum temperature, wind speed, minimum temperature, wind direction, precipitation, average relative humidity, sunshine duration, minimum relative humidity) were retained with no missing values in all features of meteorological data. To prevent overfitting while training the deep learning models, we used feature selection to remove redundant attributes. We reduced some of the meteorological variables using high correlation filtering and low variance filtering. Four variables (ie, pressure, wind speed, wind direction, sunshine duration) were discarded as they had the smallest variance in all the study areas. In total, 10 valid features (ie, six meteorological features and four types of malaria parasites) were considered in our study as shown in figure 1.

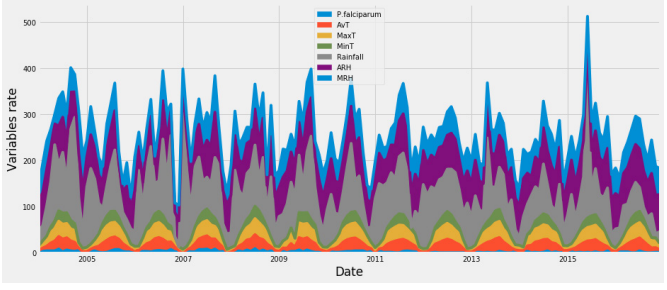

Figure 1 Guangdong climatic variables and *P. falciparum* used to train models. ARH, Average Relative Humidity; Avt, Average Temperature; MaxT, Maximum Temperature; MinT, Minimum Temperature; MRH, Minimum Relative Humidity; *P. falciparum, Plasmodium falciparum.*

Train-validation-test split

To train and evaluate the machine learning and neural network frameworks proposed in this paper, we divided the data set into the train, validation and test sets. In our experiment, 70% of the whole data set was used to train the model. We have allocated 15% of the data set for validation. The validation set was used to evaluate the model after each training epoch and ensure that the model is not overfitting the training data set. After the model has finished training, the remaining 15% of the data set was used to evaluate the model as the test set. The data was not shuffled before splitting to ensure that the validation set and test set results are more realistic. We allocated the period 1 January 2004 to 31 December 2012 to the training set and the period 1 January 2013 to 31 December 2014 is allocated to the validation set. The remaining period is allocated for the testing set.

Prediction models

This study proposes a sequence-to-sequence (Seq2Seq) prediction model based on the LSTM neural networks. The model will be used to forecast the re-emergence of malaria cases by considering the influence of meteorological factors on malaria cases in all 31 provinces of China. We compared the performance of our constructed LSTMSeq2Seq recurrent neural networks with other machine learning and deep neural networks prediction models, including XGBoost (extreme gradient boosting), GRU network and LSTM network models. Here is a brief description of our proposed Seq2Seq model as well as other employed models. These models achieved the best performance for predicting, diagnosing and controlling infectious diseases.

XGBoost model

The XGBoost is an ensemble machine learning algorithm that is flexible and easy to interpret. It provides an efficient implementation of gradient boosting machine learning model thought to be competent in the healthcare industry. A significant number of studies in public health have applied the XGBoost based framework to exploit data sources and predict infectious diseases such as dengue fever. The XGBoost model can achieve incredible performance in predicting vector-borne infectious diseases such as dengue or those caused by the West Nile virus.[29] It has been used for forecasting, prevention and early diagnosis of infectious diseases[30 31] and non-communicable diseases.[32] The hyperparameters in this gradient boosting model were tuned to optimise the XGBoost model and achieve the best performance in our study. After testing several XGBoost parameters and the number of time steps as inputs, we chose 100 trees as the number of estimators to avoid overfitting. We used the GridSearchCV method in scikit-learn to tuning the hyperparameter and a learning rate of 0.8 and a maximum depth of 8. This method greatly reduces the prediction error of our XGBoost model. We used the defined types of monthly observation *plasmodium* incidence (*P.*

falciparum, *P. vivax*, *P. malariae* and another class named other in our experiments) and climatic variables such as maximum temperature, average temperature, minimum temperature, average relative humidity, minimum relative humidity and rainfall to train the XGBoost approach and evaluate its performance on the test data set.

LSTM model

An LSTM describes a long short-term memory neural network and belongs to a class of recurrent neural networks (RNNs). RNN can process current data by using the previous data. It has effectively been used to solve problems of sequential time series such as climate modelling, web traffic prediction, financial prediction, neuroscience, intrusion detection, anomaly detection, air quality forecasting, medical monitoring. Meanwhile, RNN suffers from gradient vanishing and exploding problems when processing long-term dependencies sequences. The LSTM was developed as an intelligent recurrent neural network to specifically address the gradient vanishing problem by relying on memory cells, which have self-connections that store network temporal state, and are controlled by a set of three gates: input, output and forget. These gates and the memory cell can record information for a long time, thereby solving the problem of long-term dependencies and can predict the next time feature, which implies that it can forecast the next time step conditional on the previous values of the times series. LSTM's ability to successfully learn from data with long-range temporal dependencies makes it a natural choice for time-series predictions. This model has achieved superior performance in predicting vector-borne infectious diseases like dengue fever[33] and is one of the potential deep learning predictive models for childhood infectious diseases. It recently has been applied as one of the state-of-the-art deep neural networks in forecasting COVID-19.[34–36] We developed a two-layer LSTM model that includes 128 and 32 memory cells and uses a batch size of 32 and a diagnostic of 1000 epochs. It consists of seven input parameters for each of the four classes of *Plasmodium* species, that is, *P. falciparum*. We have the monthly observation of *P. falciparum* incidences, maximum temperature, average temperature, minimum temperature, average relative humidity, minimum relative humidity and rainfall as the input vector sequence of the same month.

GRU model

GRU is an improved recurrent neural network as a simple variant of LSTM by combining the input gate and forgetting gate into a single gate called update gate. GRU comprises of update gate and resets gate, and it can only control information inside the unit because it has no additional memory cell to keep information. Researchers have applied this framework to forecast infectious diseases such as influenza.[37] For the GRU model, we used the same hyperparameters as for LSTM models. The training data set was created using 12 months as input to our GRU model and the next month as output. The same input vector sequence as shown in figure 1 consists of seven input parameters for each of the four classes of *Plasmodium* species and six climatic variables. The six climatic variables, maximum temperature, average temperature, minimum temperature, average relative humidity, minimum relative humidity and rainfall, have been trained on the GRU model and used to test its performance.

LSTMSeq2Seq model

There are intuitively two different tasks to predict time series: understanding what has happened by looking at the known values of the past and predicting what will happen in the future. These two tasks require two different skill sets. The first is the ability to look at the past values and create an idea of the state of the system in the present. The second is the ability to use that understanding of the current state in the system to predict how the system will evolve in the future. As we mentioned earlier, LSTM predicts the next time feature, which implies that it can forecast the attribute of the next time step of input only. When we used a single LSTM cell in our model, we asked it to be capable of remembering both main events of the past and using those events to predict future values. Unlike single LSTM, we can use a Seq2Seq model with two specialised LSTM cells capable of predicting multiple time steps rather than having a single multitasking cell. Seq2Seq refers to the sequence-to-sequence architecture of the neural network fit. This architecture enables mapping between sequences of arbitrary length. As a result, Seq2Seq can perform many tasks, including language translation, image captioning and time series prediction. The Seq2Seq architecture is made up of an encoder and a decoder, as illustrated in figure 2.

LSTMSeq2Seq model consists of two major blocks: encoder LSTM cell and decoder LSTM cell. The encoder outputs the encoder vector as input to the decoder block. The decoder encodes the input vector and predicts the

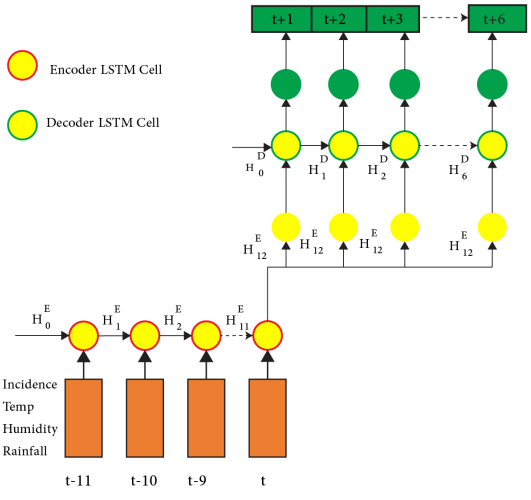

Figure 2 Long short-term memory (LSTM) sequence-to-sequence architecture.

next time step output. Subsequently, if X_t is the input of the next feature sequence, then the LSTM sequence model outputs Xt_{t+1} as the next time step feature.

The following are the formula for the encoder and decoder networks.

$$H_t^E = f\left(W^{HE} H_{t-1}^E + W^x X_t\right) \quad (1)$$

where $H^E{}_t$ represents the current hidden state at time step t, $W^{HE}{}_t$ is the appropriate weight of the old hidden state at time step t-1 and W^x represents the appropriate weight to the input vector X_t.

Equation (1) shows the result of a general sequence of the ordinary recurrent neural networks with the formula for the encoder. It is only necessary to apply an appropriate weight to the previous hidden state $H^E{}_{t-1}$ and the input vector X_t.

$$H_t^D = f\left(W^{HE} H_{t-1}^D\right) \quad (2)$$

where $H^D{}_t$ is the current decoder hidden state, we are just using the old hidden state of the input vector at some time step t-1 to compute the next one and f is some function of the parameter.

Equation (2) is a stack of numerous recurrences that forecast each output y_t at time t as a formula for the decoder. Each reiteration unit accepts a hidden state from the old unit and generates its hidden state.

The output y_t at time step t is computed using the formula (3).

$$y_t = \text{softmax}\left(W^s H_t^D\right) \quad (3)$$

y_t is the final output state at time step t computed using softmax (is used to create a probability vector which will help us determine the final output) function and its respective weight W^s.

Equation (3) calculates the output using the state hidden at the current time step with each weight W^s.

We designed an encoder that looks back into 12 months of historical data and a decoder that slide 6 months to predict, we have used t+12 months as input to the decoder as illustrated in figure 2 of our designed LSTMSeq2Seq model, the t+12 time step which is the encoder vector was used as input to the decoder and LSTM decoder cell predicts the next six steps ahead from t+1 to t+6 of malaria incidence. Apart from dropout, L1 regularisation and L2 regularisation were employed to avoid overfitting by preventing the weights of each network from being too high in the GRU, LSTM and LSTMSeq2Seq models. Each layer's high parameter values can cause the network to concentrate severely on a few features, which can lead to overfitting. Weight regularisation added a cost to the loss function of the network for large weights. As a result, the models were forced to learn only the relevant patterns in the training data.

Model validation

Using two metrics loss function scores, we evaluated the performances of our methods for predicting the re-emergence of malaria incidence based on meteorological factors. First, we used RMSE as the basis for evaluating continuous variables by measuring the average differences between predicted and observed error values.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=0}^{N}\left(y_t - \hat{y}_t\right)^2} \quad (4)$$

where y_t is the *Plasmodium* cases of observation for time t, and $\hat{y}_t$ is the number of cases predicted by the model. A lower RMSE value indicates that there is a slight difference between the predicted *Plasmodium* cases and observed ones and implicates a high prediction accuracy of the model. Second, we used mean absolute error (MAE) to assess numerically the prediction error of the sequence and calculate the average value of the errors between *Plasmodium* cases of observation for the current time step and the predicted cases.

$$\text{MAE} = \frac{1}{N}\sum_{t=0}^{N}\left|y_t - \hat{y}_t\right| \quad (5)$$

RESULTS

Comparison of LSTMSeq2Seq and candidate models

We performed all the experiments in Python (V.3.7.1) and modelled GRU, LSTM and LSTMSeq2Seq models through Tensor Flow (V.2.0.0), which is Google's application programming interface for deep learning. We also used Keras (V.2.3.1), a deep learning library used in LSTM model development (Chollet, 2015).

The main goal of this study is to develop an accurate prediction model on the re-emergence of malaria cases based on the LSTMSeq2Seq neural networks using climatic factors and malaria incidence in 31 provinces of mainland China. We applied several machine learning and deep learning predictive models to achieve our goal. We evaluated the performance of four trained models: XGBoost, GRU, LSTM and LSTMSeq2Seq methods using the above evaluation metrics (RMSE and MAE). From tables 1–4, we show the RMSE/MAE of each model, with the LSTMSeq2Seq approach showing significantly lower errors than other approaches in almost all provinces and for all four species of *Plasmodium* malaria. The prediction errors have dropped significantly in many provinces as the LSTMSeq2Seq can improve the accuracy by learning the features and fluctuations of climatic variables on malaria incidence and predicting future cases. The following figure 3 illustrates the examples of the results predicted cases for *P. falciparum*, *P. vivax*, P. *malariae* and other based on the LSTMSeq2Seq prediction model. The Y-axis represents monthly number of malaria cases for each type of *Plasmodium*. The curves show that the peak value shifts downward for *P. vivax* as the time step predicted with accurate seasonal fluctuation compared with the *P. falciparum*. We selected the provinces presented in figure 3 based on two malaria high-risk zones according to the previous studies[38 39]: the central part of China along the Huai River that consists of Henan, Hubei, Anhui and

Table 1 Comparison of model performances using the RMSE and MAE on the prediction of *Plasmodium falciparum* using climatic variables

	XGBoost		GRU		LSTM		LSTMSeq2Seq	
Province	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Anhui	0.5379	0.3102	0.3963	0.2098	0.3564	0.1873	0.1456	0.0923
Beijing	0.9426	0.7383	0.7947	0.0775	0.1705	0.0342	0.0252	0.0073
Chongqing	0.8607	0.7021	0.3854	0.1912	0.3939	0.1881	0.0553	0.0171
Fujian	0.9992	0.6264	0.7635	0.4647	0.7635	0.2016	0.6322	0.1258
Gansu	0.9761	0.8816	0.7450	0.3609	0.7464	0.2712	0.6561	0.2007
Guangdong	0.7905	0.7096	0.5614	0.4152	0.6247	0.3091	0.5284	0.2957
Guangxi	0.9842	0.6844	0.6428	0.456487	0.5329	0.3249	0.4698	0.2432
Guizhou	0.7114	0.6494	0.7059	0.5320	0.7133	0.6098	0.5603	0.3948
Hainan	0.8367	0.6704	0.6111	0.4383	0.5438	0.3222	0.4207	0.2065
Hebei	0.8229	0.6822	0.7438	0.5361	0.6683	0.3117	0.5803	0.2264
Heilongjiang	0.6183	0.5554	0.6839	0.5825	0.6242	0.5628	0.5633	0.4070
Henan	0.8239	0.6814	0.7046	0.5720	0.6533	0.5573	0.5239	0.3370
Hubei	0.8693	0.7415	0.6933	0.4469	0.5277	0.3252	0.4562	0.2156
Hunan	0.6156	0.4588	0.4025	0.2786	0.37669	0.1827	0.1787	0.0598
Inner Mongolia	0.2227	0.1507	0.1040	0.0844	0.0596	0.0361	0.0261	0.0194
Jiangsu	1.9567	1.8256	1.8880	0.9470	1.9506	1.2374	0.5005	0.3104
Jiangxi	0.7740	0.6524	0.6883	0.5059	0.6352	0.4357	0.4073	0.3237
Jilin	0.6215	0.4686	0.6204	0.4434	0.6185	0.4558	0.6095	0.4228
Liaoning	0.3949	0.2949	0.3289	0.2251	0.1213	0.0224	0.0703	0.0143
Ningxia	0.1798	0.0974	0.1609	0.0506	0.1579	0.1530	0.1500	0.0890
Qinghai	0.1870	0.0918	0.1843	0.0752	0.1829	0.0554	0.1823	0.0514
Shaanxi	0.966	0.7857	0.8323	0.6804	0.8312	0.6778	0.6731	0.4936
Shandong	0.9537	0.7626	0.7305	0.6079	0.6412	0.4879	0.4679	0.3660
Shanghai	0.6511	0.4639	0.6395	0.4242	0.5056	0.2166	0.3331	0.1080
Shanxi	0.3683	0.1744	0.1555	0.0748	0.1539	0.0626	0.1566	0.0591
Sichuan	0.7072	0.6210	0.5700	0.3088	0.5023	0.3693	0.3906	0.1235
Tianjin	0.3474	0.2332	0.3160	0.1487	0.3087	0.1504	0.2040	0.0554
Tibet	0.1494	0.0353	0.1016	0.0181	0.1017	0.0177	0.1183	0.0233
Xinjiang	0.3643	0.2157	0.2868	0.1115	0.2872	0.1367	0.2275	0.0614
Yunnan	0.9243	0.7511	0.5736	0.3699	0.6099	0.3743	0.6060	0.3783
Zhejiang	0.5508	0.2933	0.4985	0.2780	0.4404	0.1768	0.2723	0.0259

GRU, gated recurrent unit; LSTM, long short-term memory; LSTMSeq2Seq, LSTM sequence-to-sequence; MAE, mean absolute error; RMSE, root mean squared error; XGBoost, extreme gradient boosting.

Jiangsu provinces and the southwestern, southern regions which mainly comprising Guangdong, Guangxi, Hainan and Yunnan provinces. *P. vivax* was the dominant species in the first region as its climate is subtropical humid to subhumid monsoon. The LSTMSeq2Seq model achieved superior performance compared with other candidate models in most provinces with an average prediction accuracy of 87.3%. Models ranking from high performance to the lowest in the entire study are LSTMSeq2Seq, LSTM, GRU and XGBoost. LSTMSeq2Seq generates the lowest RMSE values of 0.0252, 0.0107, 0.0586 and 0.0077 for *P. falciparum*, *P. vivax*, *P. malariae* and other *plasmodia*, respectively. The LSTMSeq2Seq model reduced the mean RMSE of the predictions by 19.05% to 33.93%, 18.4% to 33.59%, 17.6% to 26.67% and by 13.28% to 21.34%, for *P. falciparum*, *P. vivax*, *P. malariae* and other *plasmodia*, respectively, as compared with other candidate models.

Since 2008 the peak value shifted downward for *P.vivax* in different regions with a significant reduction but for the *P. falciparum*, there was an increase of trends which may be due to other factors apart from climate predictors like in Guangxi province in 2013 experienced the highest incidence because of the return of Chinese labours from gold mining in Ghana. However, the increasing trends of

Table 2 Comparison of model performances using the RMSE and MAE on the prediction of *Plasmodium vivax* using climatic variables

	XGBoost		GRU		LSTM		LSTMSeq2Seq	
Province	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Anhui	0.5379	0.3102	0.3963	0.2098	0.3564	0.1873	0.1456	0.0923
Beijing	0.9426	0.7383	0.7947	0.0775	0.1705	0.0342	0.0252	0.0073
Chongqing	0.8607	0.7021	0.3854	0.1912	0.3939	0.1881	0.0553	0.0171
Fujian	0.9992	0.6264	0.7635	0.4647	0.7635	0.2016	0.6322	0.1258
Gansu	0.9761	0.8816	0.7450	0.3609	0.7464	0.2712	0.6561	0.2007
Guangdong	0.7905	0.7096	0.5614	0.4152	0.6247	0.3091	0.5284	0.2957
Guangxi	0.9842	0.6844	0.6428	0.456487	0.5329	0.3249	0.4698	0.2432
Guizhou	0.7114	0.6494	0.7059	0.5320	0.7133	0.6098	0.5603	0.3948
Hainan	0.8367	0.6704	0.6111	0.4383	0.5438	0.3222	0.4207	0.2065
Hebei	0.8229	0.6822	0.7438	0.5361	0.6683	0.3117	0.5803	0.2264
Heilongjiang	0.6183	0.5554	0.6839	0.5825	0.6242	0.5628	0.5633	0.4070
Henan	0.8239	0.6814	0.7046	0.5720	0.6533	0.5573	0.5239	0.3370
Hubei	0.8693	0.7415	0.6933	0.4469	0.5277	0.3252	0.4562	0.2156
Hunan	0.6156	0.4588	0.4025	0.2786	0.37669	0.1827	0.1787	0.0598
Inner Mongolia	0.2227	0.1507	0.1040	0.0844	0.0596	0.0361	0.0261	0.0194
Jiangsu	1.9567	1.8256	1.8880	0.9470	1.9506	1.2374	0.5005	0.3104
Jiangxi	0.7740	0.6524	0.6883	0.5059	0.6352	0.4357	0.4073	0.3237
Jilin	0.6215	0.4686	0.6204	0.4434	0.6185	0.4558	0.6095	0.4228
Liaoning	0.3949	0.2949	0.3289	0.2251	0.1213	0.0224	0.0703	0.0143
Ningxia	0.1798	0.0974	0.1609	0.0506	0.1579	0.1530	0.1500	0.0890
Qinghai	0.1870	0.0918	0.1843	0.0752	0.1829	0.0554	0.1823	0.0514
Shaanxi	0.966	0.7857	0.8323	0.6804	0.8312	0.6778	0.6731	0.4936
Shandong	0.9537	0.7626	0.7305	0.6079	0.6412	0.4879	0.4679	0.3660
Shanghai	0.6511	0.4639	0.6395	0.4242	0.5056	0.2166	0.3331	0.1080
Shanxi	0.3683	0.1744	0.1555	0.0748	0.1539	0.0626	0.1566	0.0591
Sichuan	0.7072	0.6210	0.5700	0.3088	0.5023	0.3693	0.3906	0.1235
Tianjin	0.3474	0.2332	0.3160	0.1487	0.3087	0.1504	0.2040	0.0554
Tibet	0.1494	0.0353	0.1016	0.0181	0.1017	0.0177	0.1183	0.0233
Xinjiang	0.3643	0.2157	0.2868	0.1115	0.2872	0.1367	0.2275	0.0614
Yunnan	0.2243	0.1511	0.1016	0.0699	0.1099	0.0243	0.0107	0.0083
Zhejiang	0.5508	0.2933	0.4985	0.2780	0.4404	0.1768	0.2723	0.0259

GRU, gated recurrent unit; LSTM, long short-term memory; LSTMSeq2Seq, LSTM sequence-to-sequence; MAE, mean absolute error; RMSE, root mean squared error; XGBoost, extreme gradient boosting.

P. falciparum cases in Guangdong, Hainan and Jiangsu can be predicted well by LSTMSeq2Seq with superior accuracy to traditional machine learning model and better than deep learning state-of-the-art-models employed in this study. Thus LSTMSeq2Seq can be effectively applied to the prediction of malaria re-emergence in provinces with malaria incidence.

DISCUSSION

In this study, we assessed the climatic factors that can affect the re-emergence of malaria incidence and built an advanced LSTMSeq2Seq deep neural networks model to predict the re-emergence of malaria in 31 provinces of China. We drew a comparison between the performance of the LSTMSeq2Seq model with other machine learning models applied in the study. The 2014 international panel report on climate change exposed an association between climate change and a significant increase in malaria burden.[40 41] Previous studies suggested that climatic factors are not the only cause of malaria re-emergence since other non-climatic factors are also responsible.[41] Besides climate change, malaria re-emergence is

Table 3 Comparison of model performances using the RMSE and MAE on the prediction of *Plasmodium malariae* using climatic variables

	XGBoost		GRU		LSTM		LSTMSeq2Seq	
Province	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Anhui	0.5911	0.3394	0.3767	0.1446	0.1321	0.0017	0.0586	0.0112
Beijing	0.7606	0.5225	0.5883	0.4078	0.5235	0.3623	0.1979	0.0887
Chongqing	0.5489	0.4064	0.5150	0.3611	0.39927	0.2816	0.2426	0.1707
Fujian	0.6714	0.5007	0.3003	0.2787	0.2818	0.1841	0.1551	0.0863
Gansu	0.5918	0.4138	0.4271	0.3180	0.3467	0.2137	0.2904	0.1686
Guangdong	0.6809	0.5636	0.3250	0.2898	0.3243	0.2686	0.1343	0.0856
Guangxi	0.4845	0.3817	0.3862	0.2586	0.1269	0.1059	0.1130	0.0744
Guizhou	0.4410	0.2612	0.2039	0.1495	0.1802	0.0998	0.1005	0.0694
Hainan	0.6615	0.5604	0.4981	0.3997	0.2523	0.1157	0.1791	0.1381
Hebei	0.4041	0.3601	0.3944	0.2556	0.3047	0.2418	0.2009	0.1677
Heilongjiang	0.6601	0.4212	0.4784	0.2795	0.5459	0.3318	0.5633	0.3011
Henan	0.5595	0.4855	0.1507	0.1141	0.1239	0.0846	0.0903	0.6799
Hubei	0.3672	0.3079	0.1353	0.0639	0.1869	0.0818	0.0732	0.0345
Hunan	0.4597	0.3687	0.2891	0.1960	0.2157	0.1691	0.1734	0.1159
Inner Mongolia	0.4945	0.4058	0.4142	0.3459	0.4942	0.3571	0.4672	0.3040
Jiangsu	0.5721	0.5309	0.4816	0.3630	0.4521	0.3157	0.2110	0.1850
Jiangxi	0.4434	0.3235	0.3841	0.2957	0.3329	0.2584	0.2157	0.1608
Jilin	0.4820	0.2595	0.4804	0.2540	0.4146	0.2193	0.3549	0.1024
Liaoning	0.5104	0.4233	0.4466	0.3153	0.3809	0.1781	0.2053	0.1498
Ningxia	0.4507	0.3375	0.4812	0.3101	0.4485	0.3011	0.4127	0.2923
Qinghai	0.4485	0.3041	0.3724	0.2583	0.3516	0.2433	0.2088	0.1751
Shaanxi	0.5382	0.4932	0.5257	0.4586	0.53162	0.4812	0.5158	0.4474
Shandong	0.4269	0.3949	0.4158	0.3926	0.3574	0.2148	0.2721	0.1915
Shanghai	0.5082	0.4763	0.4651	0.3680	0.3611	0.3362	0.33974	0.2777
Shanxi	0.7831	0.6217	0.6569	0.5564	0.6307	0.5466	0.6217	0.5386
Sichuan	0.4214	0.3695	0.3586	0.3238	0.3297	0.2296	0.2756	0.1269
Tianjin	0.5931	0.4835	0.5733	0.4306	0.5403	0.4294	0.4177	0.3475
Tibet	0.5952	0.3649	0.5712	0.3770	0.5891	0.3850	0.5657	0.3438
Xinjiang	0.6445	0.4381	0.4561	0.3257	0.411409	0.3052	0.3235	0.2982
Yunnan	0.5689	0.4386	0.5068	0.4156	0.4283	0.3925	0.3798	0.3452
Zhejiang	0.3723	0.2114	0.3293	0.1642	0.2832	0.1306	0.1121	0.0854

GRU, gated recurrent unit; LSTM, long short-term memory; LSTMSeq2Seq, LSTM sequence-to-sequence; MAE, mean absolute error; RMSE, root mean squared error; XGBoost, extreme gradient boosting.

affected by other global changes such as demographic shifts, increased travel and trade. Although these non-climatic factors affect malaria transmission spatiotemporally, the climatic factors facilitate the transmission by providing a suitable environment for mosquito vector activities and *Plasmodium* incubation that cause an increase in the susceptible population. Based on these findings from the previous studies, we exploit the advantages of deep learning models in handling large data sets and use them to investigate the influence of climatic factors on malaria re-emergence. Researchers have developed malaria prediction models using climate determinants and malaria incidence data in different regions. However, to the best of our knowledge, this is the first time an LSTMSeq2Seq model was employed to construct a malaria re-emergence prediction model using climate determinants and malaria incidence data in all 31 provinces of China. By comparing the performance of the proposed model with that of other candidate models, LSTMSeq2Seq has proved to have a lower prediction error value in most of the provinces for different Plasmodium species. LSTMSeq2Seq has shown excellent ability to capture trends and seasonal patterns, especially for *P. vivax* and *P. malariae*, as most of the *P. vivax* cases were

Table 4 Comparison of model performances using the root RMSE and MAE on the prediction of other *Plasmodium* species using climatic variables

	XGBoost		GRU		LSTM		LSTMSeq2Seq	
Province	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Anhui	0.4874	0.3889	0.3295	0.2963	0.3012	0.2342	0.2181	0.1605
Beijing	0.3272	0.2796	0.2591	0.1682	0.2475	0.1251	0.1578	0.0871
Chongqing	0.3696	0.2535	0.3049	0.2152	0.1639	0.1051	0.0971	0.0448
Fujian	0.5024	0.2882	0.5064	0.2697	0.46437	0.2297	0.3334	0.2209
Gansu	0.2582	0.1253	0.2045	0.0818	0.2108	0.0848	0.2059	0.0852
Guangdong	0.7559	0.5772	0.5154	0.4524	0.4236	0.3575	0.37998	0.2817
Guangxi	0.4600	0.3387	0.3313	0.2712	0.3334	0.2883	0.2566	0.1869
Guizhou	0.5307	0.3384	0.5223	0.3333	0.5250	0.3001	0.3101	0.2431
Hainan	0.5492	0.5223	0.4673	0.2379	0.3619	0.1003	0.2005	0.0802
Hebei	0.6787	0.4656	0.5882	0.4501	0.3910	0.2924	0.2667	0.1608
Heilongjiang	0.4588	0.3883	0.4101	0.3078	0.3954	0.2184	0.2111	0.1075
Henan	0.4141	0.3973	0.3692	0.2810	0.2512	0.0911	0.2357	0.0865
Hubei	0.3685	0.2202	0.2454	0.1864	0.2314	0.1635	0.1929	0.1283
Hunan	0.4476	0.3972	0.3273	0.3121	0.3924	0.2805	0.2867	0.1888
Inner Mongolia	0.3902	0.2806	0.3432	0.2482	0.3237	0.2616	0.3351	0.2139
Jiangsu	0.3968	0.2273	0.38090	0.2068	0.3137	0.1956	0.2559	0.1740
Jiangxi	0.3547	0.2902	0.3037	0.1289	0.2983	0.1258	0.2487	0.1238
Jilin	0.4449	0.4170	0.4542	0.4001	0.4342	0.3781	0.4082	0.3153
Liaoning	0.2722	0.1743	0.2479	0.1564	0.2165	0.1431	0.1356	0.0565
Ningxia	0.3748	0.2996	0.2965	0.1592	0.2636	0.1093	0.1282	0.0658
Qinghai	0.2827	0.1691	0.1358	0.0527	0.2318	0.1197	0.0691	0.0243
Shaanxi	0.3776	0.3369	0.3269	0.2107	0.2546	0.1866	0.2158	0.1319
Shandong	0.6710	0.5566	0.5630	0.4363	0.4605	0.3390	0.2611	0.1611
Shanghai	0.5067	0.3633	0.4926	0.3549	0.3935	0.2952	0.3409	0.2511
Shanxi	0.3936	0.2832	0.3801	0.2782	0.3055	0.2180	0.1224	0.0532
Sichuan	0.7541	0.5391	0.5796	0.4442	0.4232	0.3911	0.3368	0.2181
Tianjin	0.3161	0.1875	0.1076	0.0810	0.0971	0.0659	0.0930	0.0468
Tibet	0.6972	0.3431	0.46318	0.2752	0.4011	0.2112	0.3927	0.1920
Xinjiang	0.0702	0.0571	0.0455	0.0203	0.0111	0.0112	0.0073	0.0026
Yunnan	0.2590	0.2245	0.2369	0.1778	0.1832	0.1195	0.1288	0.0846
Zhejiang	0.4202	0.2507	0.2534	0.1305	0.1705	0.1176	0.1449	0.7882

GRU, gated recurrent unit; LSTM, long short-term memory; LSTMSeq2Seq, LSTM sequence-to-sequence; MAE, mean absolute error; RMSE, root mean squared error; XGBoost, extreme gradient boosting.

autochthonous and influenced by climatic factors, while *P. falciparum* cases may be imported and influenced by other global change factors. The climatic factors have proven to be effective predictors for malaria incidence and significantly affect the proposed LSTMSeq2Seq recurrent neural network models in capturing seasonal patterns and trends and predicting malaria incidence.

However, due to the fewer malaria cases in some provinces and a relatively small data set for a Seq2Seq deep neural network, GRU and XGBoost achieved lower RMSE/MAE values than the proposed method in some cases. Even so, the LSTMSeq2Seq model produced improved predictions and was better than other candidate models for each of the *Plasmodium* species in many provinces of China. However, for further improvement of malaria re-emergence prediction in China, our future research will consider climatic and non-climatic factors such as population movements, demographic shifts, changes in land use and civil unrest. By considering other potential factors that may contribute to the re-emergence of malaria incidence, we will increase the size of the data set and provide more patterns for *Plasmodium* species. We will also consider a deep learning technique known as transfer learning. This technique uses the learnt

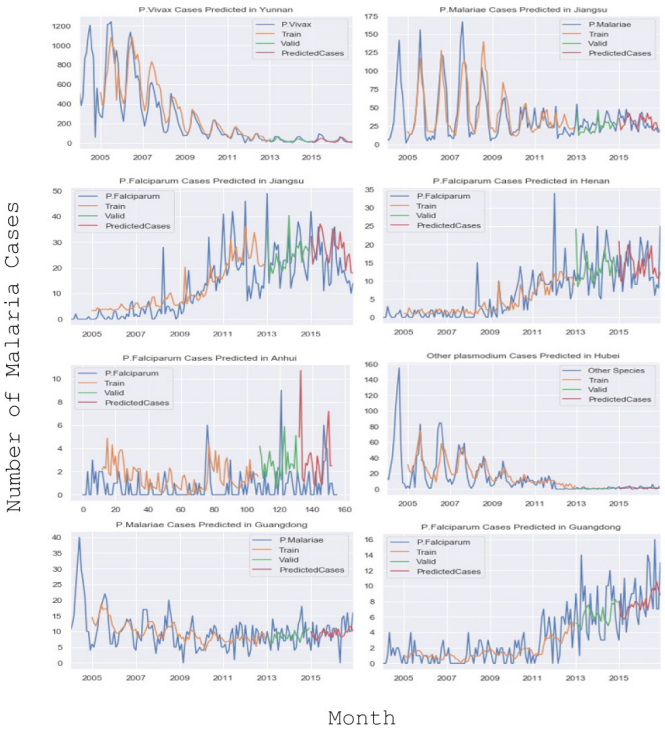

Figure 3 Predicted cases for four *Plasmodium* types using long short-term memory sequence-to-sequence model.

tusk related to the new tusk to accelerate its training and improve its predictive accuracy. It will reduce the prediction error value of the LSTMSeq2Seq in the provinces with fewer malaria cases through transfer from the previously trained model in regions with high malaria cases. Based on the LSTMSeq2Seq model, this research achieved accurate prediction of malaria cases in China, using long-term time series malaria cases and the data of climatic variables. This method might be used for the large-scale prediction of other malaria-like diseases.

There are some limitations to this study. First, the LSTMSeq2Seq takes more time for training than other employed deep learning models. To train the LSTMSeq2Seq from scratch for all 31 provinces takes 2 weeks for four types of *Plasmodium* used in our study, whereas other models take a few hours to days to train them using malaria cases and data of meteorological variables. For most cases, LSTM was seven times faster than the LSTMSeq2Seq model. However, the impact model is not significant in provinces with fewer malaria cases. Second, we could not obtain accurate predictions in some provinces by using any model in this study, probably because we failed to get other relevant potential non-climatic factors.

CONCLUSION

Malaria is still a public health burden that can be widely transmitted through the influence of many factors. To reduce this burden, it is very important to predict the re-emergence of malaria and put in place serious control measures. In this study, we investigated the influence of climatic factors in the re-emergence of malaria in mainland China by proposing an LSTMSeq2Seq model capable of effectively predicting malaria incidence using climatic factors and different types of *Plasmodium* species in all 31 provinces of China. We compared typical machine learning and other recurrent neural networks models with the performance of the LSTMSeq2Seq approach. Remarkably, the prediction performance observed in this paper indicates that LSTMSeq2Seq prediction performance outperforms the other candidate models applied in the study. Therefore, the LSTMSeq2Seq model can be effectively applied in the malaria re-emergence prediction.

Twitter Eric Kamana @kameri16

Contributors EK analysed and preprocessed the data, trained and evaluated the performance of the models, interpreted results and wrote the manuscript. JZ supervised, coordinated the design of the entire study and reviewed and edited the manuscript. JZ was the guarantor. DB collected the data used in this study. All authors have read and agreed to the submission version of the manuscript.

Funding This work was supported by the Shandong Provincial Natural Science Foundation, China (ZR2018MH037).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The research protocol was approved by the institutional review board of the Institute of Complexity Science, Qingdao University, China.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Malaria cases for all 31 provinces of mainland China were obtained at https: www.phsciencedata.cn and the meteorological data at https://data.cma.cn/en.

ORCID iD

Eric Kamana http://orcid.org/0000-0003-0829-4261

REFERENCES

1 World Health Organization. *World malaria report 2020: 20 years of global progress and challenges*, 2020.
2 World Health Organization. *Global technical strategy for malaria 2016-2030*. World Health Organization, 2015.
3 Yin J-H, Zhou S-S, Xia Z-G, *et al*. Historical patterns of malaria transmission in China. *Adv Parasitol* 2014;86:1–19.
4 Feng X, Levens J, Zhou X-N. *Protecting the gains of malaria elimination in China*, 2020: 1–3.
5 Yin J-hai, Yang M-ni, Zhou S-sen, *et al*. Changing malaria transmission and implications in China towards national malaria elimination programme between 2010 and 2012. *PLoS One* 2013;8:e74228.
6 Zhou S, Li Z, Cotter C, *et al*. Trends of imported malaria in China 2010-2014: analysis of surveillance data. *Malar J* 2016;15:39.
7 World Health Organization. *The E-2020 initiative of 21 malaria-eliminating countries: 2019 progress report. No. WHO/CDS/GMP/2019.07*. World Health Organization, 2019.
8 Andriopoulos P, Economopoulou A, Spanakos G, *et al*. A local outbreak of autochthonous Plasmodium vivax malaria in Laconia, Greece--a re-emerging infection in the southern borders of Europe? *Int J Infect Dis* 2013;17:e125–8.
9 Cohen JM, Smith DL, Cotter C, *et al*. Malaria resurgence: a systematic review and assessment of its causes. *Malar J* 2012;11:122.

10 Gao H-W, Wang L-P, Liang S, *et al*. Change in rainfall drives malaria re-emergence in Anhui Province, China. *PLoS One* 2012;7:e43686.
11 Xiang J, Hansen A, Liu Q, *et al*. Association between malaria incidence and Meteorological factors: a multi-location study in China, 2005-2012. *Epidemiol Infect* 2018;146:89–99.
12 Zhou SS, Huang F, Wang JJ, *et al*. Geographical, Meteorological and vectorial factors related to malaria re-emergence in Huang-Huai river of central China. *Malar J* 2010;9:337.
13 Huang F, Zhou S, Zhang S, *et al*. Meteorological factors-based spatio-temporal mapping and predicting malaria in central China. *Am J Trop Med Hyg* 2011;85:560–7.
14 Bi P, Tong S, Donald K, *et al*. Climatic variables and transmission of malaria: a 12-year data analysis in Shuchen County, China. *Public Health Rep* 2003;118:65–71.
15 Huang F, Zhou S, Zhang S, *et al*. Temporal correlation analysis between malaria and Meteorological factors in Motuo County, Tibet. *Malar J* 2011;10:54.
16 Xiao D, Long Y, Wang S, *et al*. Spatiotemporal distribution of malaria and the association between its epidemic and climate factors in Hainan, China. *Malar J* 2010;9:185.
17 Zhang Y, Bi P, Hiller JE. Meteorological variables and malaria in a Chinese temperate City: a twenty-year time-series data analysis. *Environ Int* 2010;36:439–45.
18 Zinszer K, Verma AD, Charland K, *et al*. A scoping review of malaria forecasting: past work and future directions. *BMJ Open* 2012;2:e001992.
19 Wang M, Wang H, Wang J, *et al*. A novel model for malaria prediction based on ensemble algorithms. *PLoS One* 2019;14:e0226910.
20 Nkiruka O, Prasad R, Clement O. Prediction of malaria incidence using climate variability and machine learning. *Inform Med Unlocked* 2021;22:100508.
21 Jiang S, Xiao R, Wang L, *et al*. Combining deep neural networks and classical time series regression models for forecasting patient flows in Hong Kong. *IEEE Access* 2019;7:118965–74.
22 Gu J, Liang L, Song H, *et al*. A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China. *Sci Rep* 2019;9:17928.
23 edLiu L, Han M, Zhou Y. Lstm recurrent neural networks for influenza trends prediction. In: *International Symposium on bioinformatics research and applications*. Cham: Springer, 2018: 259–64.
24 Mussumeci, Elisa, and Flavio Codeco Coelho. Machine-learning forecasting for dengue epidemics-Comparing LSTM, random forest and LASSO regression. *medRxiv*2020.
25 Zhang Y. ATTAIN: Attention-based Time-Aware LSTM networks for disease progression modeling. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-2019)*, Macao, China, 2019:4369–75.
26 Jia W, Wan Y, Li Y. Integrating multiple data sources and learning models to predict infectious diseases in China. *AMIA Summits on Translational Science Proceedings*, 2019:680.
27 CDC Digital Repository. China disease prevention and control center for infectious disease prevention and control. data from: Chinese center for disease control and prevention, 2019. Available: https://www.phsciencedata.cn/
28 China meteorological administration Digital Repository. China meteorological data network. data from: China meteorological data service centre, 2019. Available: https://data.cma.cn/
29 Gayle AA. Ai for early warning of seasonal infectious disease: Shapely additive explanations improves prediction of extraordinary West Nile virus events in Europe. *medRxiv*2020.
30 Hu C-A, Chen C-M, Fang Y-C, *et al*. Using a machine learning approach to predict mortality in critically ill influenza patients: a cross-sectional retrospective multicentre study in Taiwan. *BMJ Open* 2020;10:e033898.
31 Davagdorj K, Pham VH, Theera-Umpon N, *et al*. XGBoost-based framework for smoking-induced noncommunicable disease prediction. *Int J Environ Res Public Health* 2020;17:6513.
32 Xu J, Xu K, Li Z, *et al*. Forecast of dengue cases in 20 Chinese cities based on the deep learning method. *Int J Environ Res Public Health* 2020;17:453.
33 Pal R, Sekh AA, Kar S, *et al*. Neural network based country wise risk prediction of COVID-19. *Appl Sci* 2020;10:6448.
34 Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solitons Fractals* 2020;140:110212.
35 Zhang G, Liu X. Prediction and control of COVID-19 infection based on a hybrid intelligent model. *medRxiv*2020.
36 Venna SR, Tavanaei A, Gottumukkala RN, *et al*. A novel data-driven model for real-time influenza forecasting. *IEEE Access* 2018;7:7691–701.
37 Reiner RC, Geary M, Atkinson PM, *et al*. Seasonality of Plasmodium falciparum transmission: a systematic review. *Malar J* 2015;14:343.
38 Song Y, Ge Y, Wang J, *et al*. Spatial distribution estimation of malaria in northern China and its scenarios in 2020, 2030, 2040 and 2050. *Malar J* 2016;15:345.
39 Smith K, Woodward A, Campbell-Lendrum D. Human health: impacts, adaptation, and co-benefits. In: *Climate change 2014: impacts, adaptation, and vulnerability. Part A: global and sectoral aspects. contribution of working group II to the fifth assessment report of the Intergovernmental panel on climate change*. Cambridge University Press, 2014: 709–54.
40 Caminade C, Kovats S, Rocklov J, *et al*. Impact of climate change on global malaria distribution. *Proc Natl Acad Sci U S A* 2014;111:3286–91.
41 Alim M, Ye G-H, Guan P, *et al*. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. *BMJ Open* 2020;10:e039676.