This is an Accepted Manuscript for Epidemiology & Infection. Subject to change during the editing and production process.
DOI: 10.1017/S0950268824000694

Analysis and forecasting of syphilis trends in mainland China based on hybrid time series models

ZhenDe Wang[1] ChunXiao Yang[1] ShengKui Zhang[2] YongBin Wang[3] Zhen Xu[4]* ZiJian Feng[5]*

1 School of Public Health, Shandong Second University, Weifang, China.
2 School of Basic Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China.
3 School of Public Health, Xinxiang Medical University, Xinxiang, China.
4 Chinese Center for Disease Control and Prevention, Beijing, China.
5 Chinese Preventive Medicine Association, Beijing, China.
*Corresponding author

Abstract

Syphilis remains a serious public health problem in mainland China that requires attention, modeling to describe and predict its prevalence patterns can help the government to develop more scientific interventions. Time series (TS) data of the syphilis incidence from January 2004 to November 2023 was obtained from the website of the Bureau for Disease Control and Prevention of China National Health Commission. The seasonal autoregressive integrated moving average (SARIMA) model, long short-term memory network (LSTM) model, hybrid SARIMA-LSTM model, and hybrid SARIMA-nonlinear auto-regressive models with exogenous inputs (SARIMA-NARX) model were used to simulate the data respectively, the model performance was evaluated by calculating the R^2, median absolute deviation (MAD), mean absolute percentage error (MAPE), root mean square error (RMSE), and mean absolute error (MAE) of the train and test sets of the models. The SARIMA-NARX model predicts better than the other three models, despite its R^2 value is 3.73% lower than the SARIMA-LSTM model. Compared to the SARIMA, LSTM, and SARIMA-LSTM models, the MAD value of the SARIMA-NARX model decreases by 352.69%, 4.98%, and 3.73%, respectively. The MAPE value decreases by 73.7%, 23.46%, and 13.06%, respectively. The RMSE value decreases by 68.02%, 26.68%, and 23.78%, respectively. The MAE value decreases by 70.90%, 23.00%, and 21.80%, respectively. The hybrid SARIMA-NARX and SARIMA-LSTM methods predict syphilis cases more accurately than the basic SARIMA and LSTM methods, so that can be used for

governments to rationally allocate health resources and develop long-term syphilis prevention and control programs. In addition, the predicted cases still maintain a fairly high level of incidence, so there is an urgent need to develop more comprehensive prevention strategies.

Keywords: Syphilis, Modeling, SARIMA, LSTM, NARX

1 Introduction

Syphilis, an infectious disease caused by the bacterium Treponema pallidum, is a preventable and treatable condition predominantly transmitted through sexual contact, including oral, vaginal, and anal intercourse. Additionally, vertical transmission can occur from mother to fetus during pregnancy, and less commonly, the disease may be spread through blood transfusion. The clinical presentation of syphilis can be asymptomatic[1], rendering the identification of infected individuals challenging in the absence of serological screening. Consequently, routine testing is imperative for the detection of syphilis, particularly in populations at increased risk for sexually transmitted disease (STDs)[2, 3]. Syphilis can result in the development of genital ulcers, pain, and inflammation. Untreated, the disease can advance to affect various organs and systems. Advanced syphilis can lead to detrimental effects on the heart, major blood vessels, central nervous system, and skeletal structure, resulting in a myriad of complications such as heart valve abnormalities, meningitis, stroke, optic nerve damage, and bone degeneration.

In recent decades, the number of syphilis cases has been increasing[4]. An estimated 5.7-6 million new cases are detected annually worldwide among individuals aged 15-49 years[5]. Between 2016 and 2023, the annual reported incidence rate of congenital syphilis ranged from 700,000 to 1.5 million cases per year[6], the case-fatality rate (CFR) among offspring of pregnant women with syphilis was 31%[7]. The reported incidence of syphilis in China escalated from 4.50 per 100,000 in 2003 to 34.04 per 100,000 in 2021[8], the mortality rate from syphilis was recorded at 0.002 per 100,000 individuals[9]. Syphilis has become the highest number of reported incidences of all STDs in mainland China[10]. As a serious public health problem, it has attracted great attention from the national health authorities. A prerequisite for policymakers to develop policies is to make scientific forecasts of disease trends. Many approaches are available for modeling and forecasting TS data. The most widely used TS model is the autoregressive integrated moving average (ARIMA) model[11, 12], a prerequisite for the applicability of ARIMA models is the requirement that the TS data should be stable, so ARIMA models tend to be poor fits for nonlinear data, the real TS data tend to be more complex, containing both nonlinear and linear components. For the nonlinear component of the model, Machine Learning (ML) is a more applicable approach. ML is a field that focuses on the learning aspect of Artificial Intelligence (AI) by developing algorithms that best represent a set of data[13]. Originally inspired by neurobiology, deep neural network models have become a powerful tool of ML and artificial intelligence. They can approximate functions and dynamics by learning from examples[14]. Artificial Neural Networks (ANN) are an important part of ML, which

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

is an autonomous computational project designed by imitating the structure of biological neurons. According to the different topologies of neural networks, artificial neural networks can be divided into feed-forward, feed-back, and recurrent neural networks (RNNs), among which the RNN models, such as the long short-term memory network (LSTM) models, the nonlinear auto-regressive models with exogenous inputs (NARX) have unique advantages in processing TS data, this is because the structure of the RNN models determines that the output at moment *t* is not only related to the input at moment *t* but related to the output at moment *t*-1. LSTM, leveraging its gating mechanism and memory unit, is capable of capturing prolonged dependencies within sequential data and conducting contextual modeling, thereby exhibiting remarkable efficacy in the modeling and prediction of sequential data[15]. The NARX neural network is a dynamic neural network that incorporates delay and feedback mechanisms, thereby enhancing its ability to memorize historical data. It is suitable for simulating and predicting nonlinear time-series data in multiple domains[16]. Compared to other data-driven models, the NARX model demonstrates strong problem-solving capabilities, fast convergence speed, and high prediction accuracy when handling seasonal time series forecasting[17]. Since ARIMA models have a good ability to fit the linear component of TS data, many researchers combine the ARIMA models and RNN models into hybrid ARIMA-RNN models to predict the convergence of the TS data, and the prediction results of the hybrid models are often better than those of single ARIMA or RNN models[18-23]. In this study, we simulated the TS data of syphilis incidence using a single ARIMA model, LSTM model, hybrid ARIMA-LSTM model, and hybrid ARIMA-NARX model, respectively, both models were used to make predictions with a period of 12. The fit and prediction indicators were calculated separately to evaluate the performance of these models.

2 Methods

2.1 Data collection

The monthly number of newly reported cases of syphilis from January 2004 to November 2023 was obtained from the website of the Bureau for Disease Control and Prevention of China National Health Commission (*http://www.nhc.gov.cn/jkj/new_index.shtml*) by searching for the Chinese translation of the keyword "*Overview of National Notifiable Infectious Diseases Epidemic*" in the website's search box. The monthly reports were simultaneously published on the China National Knowledge Infrastructure (CNKI, https://www.cnki.net/). The case information of notifiable infectious diseases was timely reported from local hospitals and community health service centers throughout the country and was reviewed and confirmed by local Centers for Disease Control and Prevention (*CDC*) after confirmatory tests[24]. A total of N = 239 observations were included in the study.

2.2 TS decomposition

TS decomposition is a technique to break down a time series into its underlying components. which is expressed as $Yt = T_t + S_t + I_t$, where T_t, S_t, and I_t denote the trend, seasonal component, and a stochastic irregular component, respectively. We performed

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

the Mann-Kendall (M-K) test for the trend .As the sample data is monthly, we confirmed the T_t by using a smooth weighted 13-term moving average filter given by:

$$T_t = \sum_{j=-q}^{q} k_j \, y_{t+j}$$

$q = 6$ for monthly data, $q < t < N - q$, $k_j = 1/4q$ for $j = \pm q$, and $k_j = 1/2q$ otherwise. After the transformation of TS, the first and last q observations were lost, so we repeated the first and last smoothed values q times.

Let n_t be the total number of observations made in period t, the stable seasonal filter is given by

$$\widetilde{s_t} = \frac{1}{n_t} \sum_{j=1}^{\left(\frac{N}{s}\right)-1} x_{t+js}$$

$$\bar{s} = \frac{1}{S} \sum_{t=1}^{s} \widetilde{s_t}$$

$$\widehat{S_t} = \widetilde{s_t} - \bar{s}$$

For $s = 12$, $t = 1,...,s$. Using $\widehat{s_t}$ to constrain the seasonality component to fluctuate around *zero*.

2.3 Modeling of SARIMA

2.3.1 Mathematical equations of the SARIMA model

The SARIMA model is always expressed as SARIMA *(p, d, q) (P, D, Q)$_s$*, where *p, d, and q* represent non-seasonal components, and *P, D, and Q* represent seasonal components. *p* and *P* are lags of non-seasonal and seasonal autoregressive, respectively. *d and D* are degrees of non-seasonal and seasonal differencing, respectively. *q* and *Q* are lags of the non-seasonal and seasonal moving average, respectively, and *s* denote the periodicity.

The polynomial of SARIMA *(p, d, q) (P, D, Q)$_s$* model can be expressed as

$$\varphi(L)\ \Phi(L)\ \Delta^d\ \Delta_s^D\ y_t = \theta(L)\ \Theta(L)\ \varepsilon_t$$

$$L^i y_t = y_{t-i}$$

$$\Delta^d = (1 - L)^d$$

$$\Delta_s = (1 - L^s)$$

$$\varphi(L) = 1 - \varphi_1 L - \ldots - \varphi_p L^p$$

$$\Phi(L) = 1 - \Phi_s L - \ldots - \phi_{Ps} L^{Ps}$$

$$\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q$$

$$\Theta(L) = 1 + \Theta_s L + \ldots + \Theta_{Qs} L^{Qs}$$

Where ε_t denotes a sequence of uncorrelated random variables from a defined probability distribution with a mean zero.

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

2.3.2 Constructing the SARIMA model

Initially, we conducted an Augmented Dickey-Fuller (ADF) test to determine the lags of seasonal and non-seasonal differencing required to achieve data stationarity. Subsequently, we established a range for the p, q, P, Q parameters, from 0 to 4. We computed the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for each permutation of the SARIMA model. The model exhibiting the lowest combined sum of AIC and BIC was selected as the optimal SARIMA model, the parameters of the model were estimated by the maximum likelihood approach. We conducted a Ljung-Box Q-test, along with the ACF and PACF plots on the residuals to check the autocorrelation. Besides, we performed normality diagnostics by plotting the histogram of standard residuals and the Quantile-Quantile (QQ) plot of residuals. The TS was devided into a training set (the first 227 observations) for modeling and a test set(the last 12 observations) for predicting. Finally, the simulation performance of the training set and the test set were calculated separately.

2.4 Constructing a single LSTM model

2.4.1 Structure and equation of LSTM neural network

The LSTM network is a kind of RNN consisting of a sequence input layer, an LSTM layer, and an output layer. The sequence input layer inputs TS data into the network, and the LSTM layer learns long-term dependencies between time steps of sequence data. Different from the traditional RNN, there is a cell state in the LSTM layer, which can effectively keep the long-term information learned from the previous time steps and solve the problem of gradient disappearance. At each time step, the layer adds information to or removes information from the cell state, all these updates are controlled by gates. There are three kinds of gates in the LSTM layer, input gate (i), forget gate (f), and output gate (o). **Figure 1** illustrates the flow of data at time step t and shows how the gates forget, update, and output the cell and hidden states.

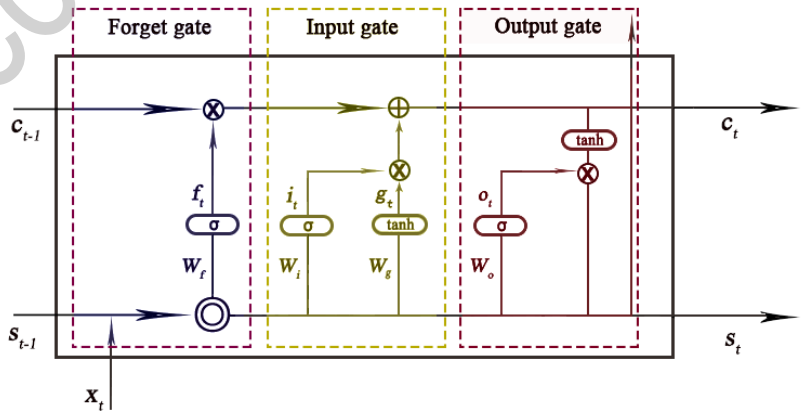


Figure 1 The cell structure of the LSTM network

Note: The arrow indicates the data flow, where x, s, c, f, i, g, and o denote the input, output, cell state, forget gate, input gate, cell candidate, and output gate in time step t, respectively. σ and *tanh* denote the *sigmoid* activation function and the hyperbolic tangent function, which maps the data to (0,1) and (-1,1),

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

respectively. ⊗ are vector operators which represent element-wise multiplication and element-wise addition, respectively.

The following formulas describe the operation of the data in the LSTM layers at time step *t*.

$$f_t = \sigma\left(W_f \cdot [S_{t-1},\ X_t\ ] + b_f\right)$$

$$i_t = \sigma(W_i \cdot [S_{t-1}, X_t\ ] + b_i)$$

$$g_t = tanh\left(W_g \cdot [S_{t-1},\ X_t\ ] + b_g\right)$$

$$o_t = \sigma(\ W_o \cdot [S_{t-1}, X_t\ ] + b_o)$$

$$C_t = f_t \otimes C_{t-1} +\ \ i_t \otimes g_t$$

$$S_t = O_t \otimes tanh(C_t)$$

Where *W*, *b* denote the matrices of input weight and bias, respectively.

2.4.2 Training and simulation of the LSTM model

It is necessary to standardize the data before modeling to eliminate the effects of abnormal values and improve the speed of model convergence. A *z*-score method was used to normalize the TS data, which was given by $TS^* = (TS - \mu) / \sigma$, where μ and σ denote the mean and standard deviation of the TS data. To prevent the gradients from exploding, we set the gradient threshold of the network to 1. The loss function is an important basis for adjusting parameters, it reflects the difference between the original and predicted values during the training, if the loss function decreases too slowly at the initial stage of training, it may mean that the number of hidden neurons or the value of the learning rate is set too small. We used the *Adam* solver to update the network parameters by taking small steps in the direction of the negative gradient of the loss function. The solvers update the parameters using a subset of the data at each step.

The varying quantities of hidden neurons, iterations, and learning rates can all influence the simulation performance of the model. Therefore, a commonly used method for adjusting parameters involves fixing the learning rate and iteration count, and then conducting model training with different numbers of neurons[25]. Presently, there is no mature theoretical evidence for determining the optimal number of neurons. Consequently, the majority of studies rely on trial and error[26]. In our study, we set the initial learning rate to 0.005, which is the median value of recommended[27], the max iterations wae set to 500, and the number of hidden neurons in increments of 10, ranging from 10 to 200. To automatically drop the learning rate during training, using a piecewise learn rate schedule, multiply the initial learning rate by a drop factor of 0.2 after half of the maximum iterations. To mitigate the risk of overfitting, we incorporated L2 regularization and implemented dropout layers within the model. We then calculated the goodness-of-fit for the test set under different numbers of neurons, using the principle of minimizing the RMSE to determine the best-fitting LSTM model. The first

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

226 data of the training set were set as input of the model, and the data shifted to one-time step were set as output of the model. A 12-step forward prediction was then performed using the trained LSTM model. Finally, the simulation performance of the training set and the test set were calculated separately.

2.4.3 Constructing a hybrid SARIMA-LSTM model

The thought of constructing the hybrid model is to express the link between the output of SARIMA and original observations (i.e., the residuals of the SARIMA model) by an LSTM model, for the residuals of the SARIMA model containing the time informations and random fluctuations. Although the SARIMA model effectively captures the linear components of the TS data, its capability to capture non-linear features is comparatively limited when compared to the LSTM model. Consequently, the residuals of the SARIMA model may contain unutilized random fluctuation information from the original data. One of the key advantages of the LSTM model lies in its capacity to model stochastic data. Leveraging this capability, we employed the LSTM approach to re-model the residuals derived from the SARIMA model. Subsequently, we integrated the LSTM model's output with that of the SARIMA model to obtain the hybrid SARIMA-LSTM model's output. The modeling and prediction processes remained consistent with those of the single LSTM model. In our study, we set the initial learning rate to 0.005 , the max iterations wae set to 500, and the number of hidden neurons ranging from 10 to 200 in increments of 10. To automatically drop the learning rate during training, using a piecewise learn rate schedule, multiply the initial learning rate by a drop factor of 0.2 after half of the maximum iterations. To mitigate the risk of overfitting, we incorporated L2 regularization and implemented dropout layers within the model. We then calculated the goodness-of-fit for the test set under different numbers of neurons, using the principle of minimizing the RMSE to determine the best-fitting SARIMA-LSTM model. Subsequently, a 12-step forward prediction was executed using the trained SARIMA-LSTM model, and the simulation performance of the training and test sets was evaluated separately.

2.4.4 Constructing a hybrid SARIMA-NARX model

The NARX network is a powerful neural network architecture specifically designed for modeling and predicting time series data by considering both the autoregressive relationship within the time series and the influence of exogenous inputs.

The NARX network consists of two main components: the autoregressive (AR) part and the exogenous (X) part. The AR part captures the relationship between past values of the time series itself, while the X part captures the influence of the exogenous inputs on the time series. The X part can be implemented as a separate input layer or concatenated with the AR inputs.

During training, the NARX network is fed with historical data, including both the time series values and the corresponding exogenous inputs. The network learns to predict the future values of the time series based on its past values and the exogenous inputs. The training process involves adjusting the network's weights and biases to minimize prediction errors.

The defining equation for the NARX model is

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

$$y(t) = f\left(y(t-1), y(t-2), \ldots, y\left(t-n_y\right), u(t-1), u(t-2), \ldots, u(t-n_u)\right)$$

Where *f* represents a function that relies on the structure and connection weights of the NARX model, *y* refers to the sample TS data in a lagged period *d*, and *u* refers to the input series containing the time factor and the projections of the SARIMA model, y is the simulation values by the hybrid SARIMA-NARX model at time step *t*. Before modeling, we need to define the structure of the model. In this model, the simulated series of the SARIMA model was treated as the input, while the corresponding reported cases of syphilis were regarded as the output. Subsequently, we randomly divided the data into a training set, a validation set, and a test set in the ratio of 80%, 10%, and 10% respectively[28]. Since the delays of the input and the number of hidden neurons have an impact on the performance of the model, we constructed multiple open-loop (series-parallel) architectures containing different number of hidden neurons (experimented from 2 to 40) for training the networks separately, using the Levenberg-Marquardt algorithm for updating weights during training. We evaluated the goodness-of fit of models under different numbers of neurons, and calculated the RMSE for both the training and test sets. The SARIMA-NARX model with the smallest RMSE value on the test set was chosen as the best-fitting model.Finally, the trained open-loop network was transformed into a closed-loop (parallel) architecture to make a 12-step-ahead forecast and the goodness-of-fit for the training and test sets were calculated separately.

2.5 Goodness-of-fit checks of models

The R^2, MAD, RMSE, MAPE, and MAE of train set and test set were used as indicators for evaluating the simulation and prediction performance of the models mentioned above, which were given by

$$R^2 = 1 - \frac{\sum_t^N (x_t - y_t)^2}{\sum_t^N (x_t - \bar{x_t})^2}$$

$$MAD = \text{median}(|x_t - y_t|)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_t - y_t)^2}$$

$$MAPE = \frac{1}{N}\left(\sum_{t=1}^{N}\frac{|x_t - y_t|}{x_t}\right)$$

$$MAE = \frac{1}{N}\left(\sum_{t=1}^{N}|x_t - y_t|\right)$$

Where x_t and y_t denote the original series and fitting series, respectively.

2.6 Software and significant level

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

MATLAB 2023a (MathWorks Corporation, USA) was used to perform the models involved in the study, and Microsoft Office 2021 (Microsoft Corporation, USA) for data collection and processing. A two-sided $p<0.05$ was considered statistically significant.

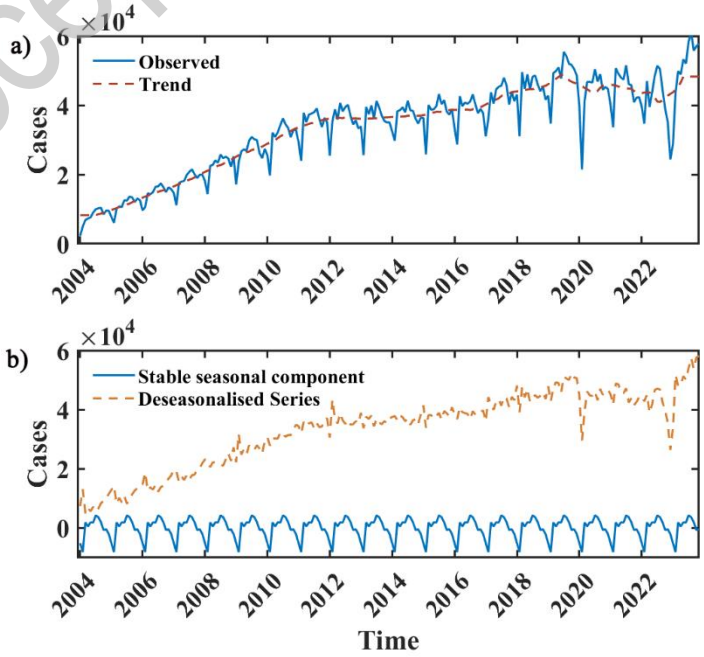
2.7 Ethical review

The study protocol and utilization of syphilis incidence data were obtained from the Bureau for Disease Control and Prevention of China National Health Commission and no ethical issues were identified. Therefore, an ethical statement was not necessary because the data are public access data.

3 Results

3.1 Trends and seasonality of the sample data

The monthly average number of reported cases for the last three years was 45,735, which is 26.74% higher than the average for the whole period. The peak number of reported incidents was 61,068 in November 2023. The M-K test results indicate an overall increasing trend in the data ($z=17.11$, $p<0.05$). As shown in **Figure 2.a**, the number of syphilis infections trended upward from 2004 until Quarter 3 of the year 2019, monthly reported cases increased at a higher rate from 2004 to 2012 than from 2013 to 2019 and has leveled off since then, however, no significant downward trend has been observed since the year 2019 ($z=0.31$, $p>0.05$). The decomposition results of the data showed a periodicity of 12 in the TS data, with a peek number of incidences in July, and less prevalence in winter and spring than in summer and autumn (**Figure 2.b**).



https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

Figure 2 Monthly reported cases of syphilis from January 2004 to November 2023 and the decomposition of the TS data

Note: In figure 2.a, the blue curve depicts the monthly reported incidences of syphilis, while the red curve illustrates the long-term trend. Meanwhile, in figure 2.b, the blue curve represents the stable seasonal component exhibiting a periodicity of twelve months, and the yellow dash curve portrays the time series post seasonal component extraction.

3.2 The best-fitting SARIMA model

The ADF test results indicate that the TS data achieves stationarity subsequent to the implementation of both a first-order differencing and a first-order seasonal differencing (t=-7.22, p<0.01), therefore d and D of the SARIMA model should be set to 1. In the process of model selection, the SARIMA(4, 1, 0)(4, 1, 0)$_{12}$ emerged as the superior model, as evidenced by the minimization of the combined AIC and BIC values (AIC=4072.3, BIC=4102.6) relative to competing models. Consequently, this SARIMA model is adjudged to be the optimal fit for the TS data under study. The best-fitting SARIMA model can be expressed as a polynomial of

$$(1 + 0.51L + 0.20L^2 + 0.16L^3 + 0.11L^4)(1 + 0.78L^{12} + 0.76L^{24} + 0.66L^{36} + 0.69L^{48})(1 - L)(1 - L^{12})y_t = \varepsilon_t.$$

We performed the autoregression and normality diagnostics on the residual series, and the result of Ljung-Box Q-Test test showed that there was no autocorrelation in the residuals (χ^2=17.81, p=0.59), and the residual ACF and PACF plots showed that most of the residuals were within the ±2 times standard deviation interval, which indicated that the fitting was successful. The histogram of the standardized residual distribution and the QQ plot of the residuals indicated that the standardized residuals showed an almost symmetrical distribution with zero as the boundary, and the frequency of the standardized residuals in the ±2 interval accounted for more than 80% of all, which can therefore be regarded as a normal distribution(**Figure 3**).

A 12-time-step prediction was performed using the SARIMA(4, 1, 0)(4, 1, 0)$_{12}$ model, The fitting and predicting efficacy of the model was calculated separately, which are shown in **Table 1**.

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

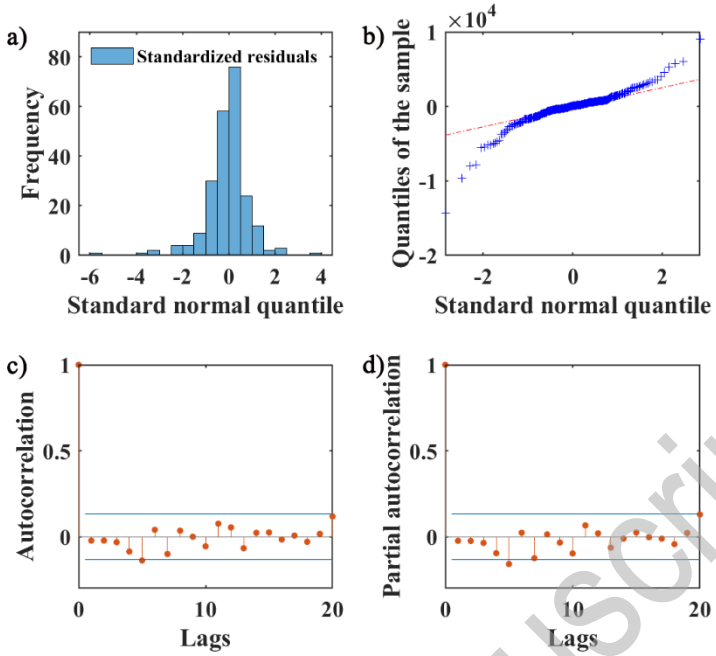


Figure 3 SARIMA model residuals normality and autocorrelation diagnostics

Note: Figure 3.a shows the frequency distribution of standardized residuals using a histogram. Figure 3.b shows the QQ plots of residuals of the SARIMA model, the red dashed line represents the standard normal distribution. Figure 3.c and Figure 3.d are the ACF and PACF of residuals, respectively. The stem plots represent the values of ACF and PACF at different lags, and the blue lines indicate the ±2 times standard deviation interval.

3.3 The best-performing LSTM model and SARIMA-LSTM model

Upon establishing a constant number of maximun iterations and an initial learning rate, we systematically trained a series of LSTM neural network architectures differentiated by the quantity of hidden neurons embedded within them. These architectures were then employed to execute simulations on both the train and test datasets. Empirical evidence suggests that the architecture containing precisely 130 hidden neurons yielded superior predictive capabilities, as quantified by the RMSE on the test set, which registered a value of 6381(**Figure 4.a**). Leveraging the LSTM model that demonstrated optimal performance characteristics, we engaged in an iterative process of retraining and forecasting. The conclusive model's fitness was quantitatively assessed, with the results systematically tabulated in **Table 1**, and the corresponding simulation outputs alongside the residual diagnostics are illustrated in **Figure 6**.The residuals of the single SARIMA model are modeled quadratically using the LSTM approach, and then the output of the LSTM is summed with the fitted values of the SARIMA model to obtain the output of the hybrid SARIMA-LSTM model. Similar to the process of determining the structure of the LSTM model, we conducted training on LSTM models with varying numbers of hidden neurons, while maintaining fixed iterations and initial learning rates. Subsequently, all models were used to simulate the training and test sets. The results indicate that the model exhibits the best predictive performance when the number of hidden neurons is 170 (the RMSE value on the test set is 6114, as shown in **Figure 4.b**).

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

The optimal SARIMA-LSTM model obtained was subjected to multiple rounds of training and prediction. The final goodness of fit results for the model are presented in **Table 1**, while the simulation results and residuals are depicted in **Figure 6**.

3.4 The best-performing SARIMA-NARX model

After several trials, we found that the best prediction performance was achieved when thenumber of hidden neurons was 30, with the minimal RMSE value(5701) of the test set (**Figure 4.c**). So the structure of the SARIMA-NARX model was determined. Then an open-loop network was built. Before training, we divided the training set data into three parts, 80% for training, 10% for validation, and 10% for testing. At the 13th iteration of the model, the MSE value of the validation set reached its minimum and began to rise(**Figure 5.a**). The output of the model training process and the errors are shown in **Figure 5**. After the training was completed, we converted the network into a closed-loop to perform the prediction for the forward 12 steps, and the goodness-of-fit was calculated for the training set and test set respectively (**Table 1**).

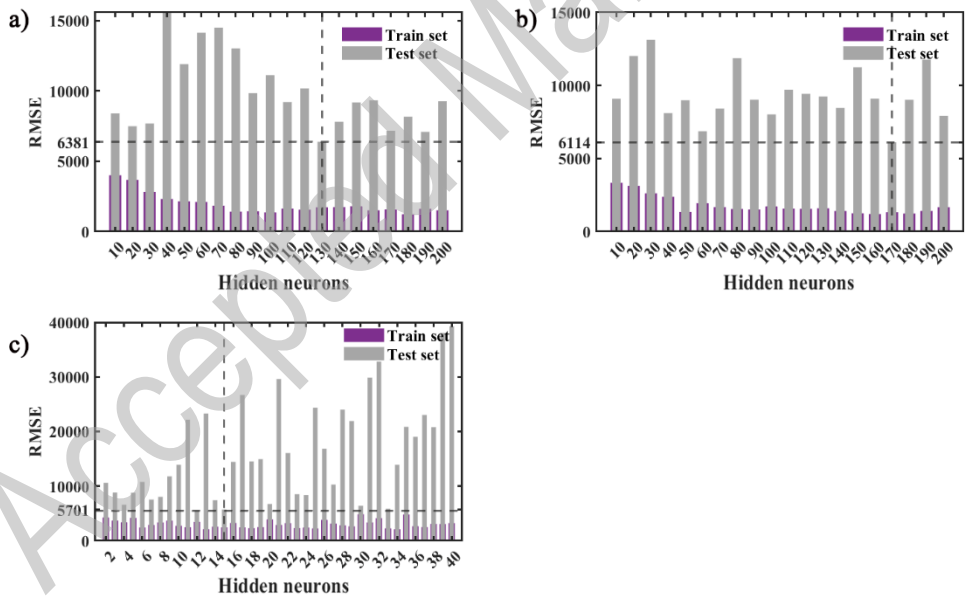


Figure 4 Fitting and predicting the performance of LSTM (Figure 4.a), SARIMA-LSTM(Figure 4.b) and SARIMA-NARX (Figure 4.c) models with different structures

Note: The dark blue and grey bars represent the RMSE values of the training and test sets, respectively.

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

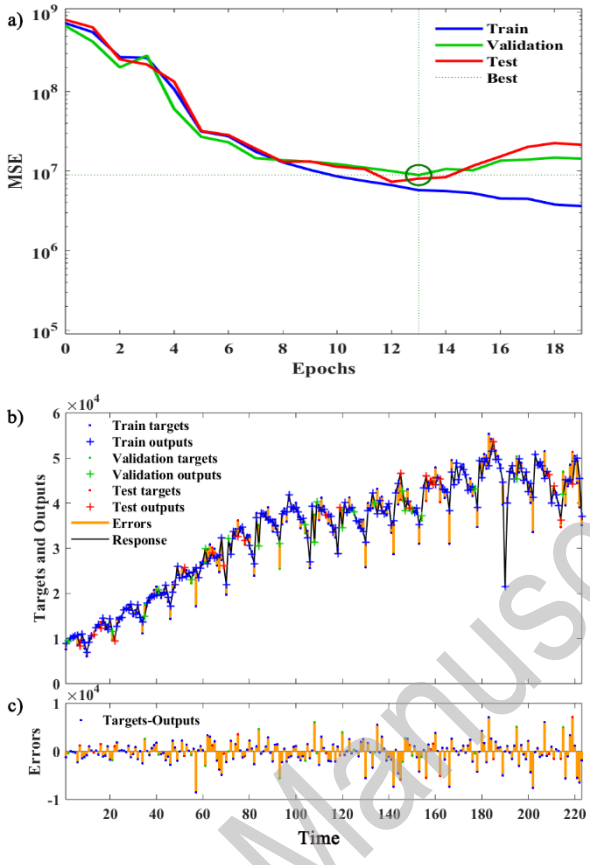


Figure 5 The simulation process for the training set by the SARIMA-NARX model

Note: Figure 5.a represents the variation of MSE for the training, validation, and test sets during the iteration process. Figure 5.b illustrates the error between the output values of each component data and the target values, while figure 5.c provides a detailed display of the error magnitude. The blue, yellow, and red dots indicate the target values of the training set, validation set, and test set after simulation using the SARIMA-NARX model, and the blue, yellow, and red crosses denote the outputs of the training set, validation set, and test set, and the yellow stem denotes the error of fitting.

3.5 Comparison of the fitting and predicting power of SARIMA, LSTM, SARIMA-LSTM, and SARIMA-NARX models

In conclusion, all four models exhibit high goodness-of-fit in the training set, with R-squared values exceeding 95%. Considering the comprehensive fit indices, the ranking of the goodness-of-fit from best to worst is SARIMA-LSTM model, LSTM model, SARIMA model, and SARIMA-NARX model. In terms of predictive performance, the SARIMA model performs the worst, almost unable to accurately predict the future epidemic trends. The SARIMA-NARX model outperforms the other models, despite its R^2 value being slightly lower than the SARIMA-LSTM model. Its MAD value decreases by 352.69%, 4.98%, and 3.73% compared to SARIMA, LSTM, and SARIMA-LSTM models, respectively. Its MAPE value decreases by 73.7%, 23.46%, and 13.06% compared to SARIMA, LSTM, and SARIMA-LSTM models, respectively. The RMSE value decreases by 68.02%, 26.68%, and 23.78% compared to SARIMA,

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

LSTM, and SARIMA-LSTM models, respectively. The MAE decreases by 70.90%, 23.00%, and 21.80% compared to SARIMA, LSTM, and SARIMA-LSTM models, respectively. Compared to the LSTM model, the SARIMA-LSTM model's MAD, MAPE, RMSE, and MAE values decrease by 11.97%, 7.47%, 3.81%, and 1.53%, respectively. From the fitting curves of the four models, it can be observed that the LSTM, SARIMA-LSTM, and SARIMA-NARX models can all accurately predict future disease trends. Among them, the predictions from the LSTM and SARIMA-LSTM models are more similar. The predictive errors of the three models reach their maximum values in July and August 2023, and the predicted values are consistently lower than the actual data(**Figure 6**).

Table 1 Evaluation of goodness-of-fit of SARIMA, LSTM, SARIMA-LSTM, and SARIMA-NARX models

Models	Fitting power					Predicting power				
	R^2	MAD	MAPE	RMSE	MAE	R^2	MAD	MAPE	RMSE	MAE
SARIMA	96.54%	869.64	5.13%	2363.59	1504.26	18.99%	13643.14	28.72%	14130.01	13522.92
LSTM	97.81%	977.18	4.84%	1856.21	1371.02	90.47%	4687.71	12.88%	6163.92	5110.33
SARIMA-LSTM	98.43%	820.38	3.94%	1569.46	1146.40	89.29%	4126.54	11.92%	5929.18	5032.13
SARIMA-NARX	95.82%	1239.12	6.11%	2508.07	1805.93	85.96%	3587.78	8.75%	4519.34	3935.07
SARIMA VS. LSTM	-1.31%	-12.37%	5.57%	21.47%	8.86%	-376.41%	65.64%	55.15%	56.38%	62.21%
SARIMA VS. SARIMA-LSTM	-1.96%	5.67%	23.25%	33.60%	23.79%	-370.24%	69.75%	58.50%	58.04%	62.79%
SARIMA VS. SARIMA-NARX	0.74%	-42.49%	-19.04%	-6.11%	-20.05%	-352.69%	73.70%	69.55%	68.02%	70.90%
LSTM VS. SARIMA-LSTM	-0.64%	16.05%	18.73%	15.45%	16.38%	1.29%	11.97%	7.47%	3.81%	1.53%
LSTM VS. SARIMA-NARX	2.03%	-26.81%	-26.06%	-35.12%	-31.72%	4.98%	23.46%	32.11%	26.68%	23.00%
SARIMA-LSTM VS. SARIMA-NARX	2.65%	-51.04%	-55.10%	-59.80%	-57.53%	3.73%	13.06%	26.63%	23.78%	21.80%

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

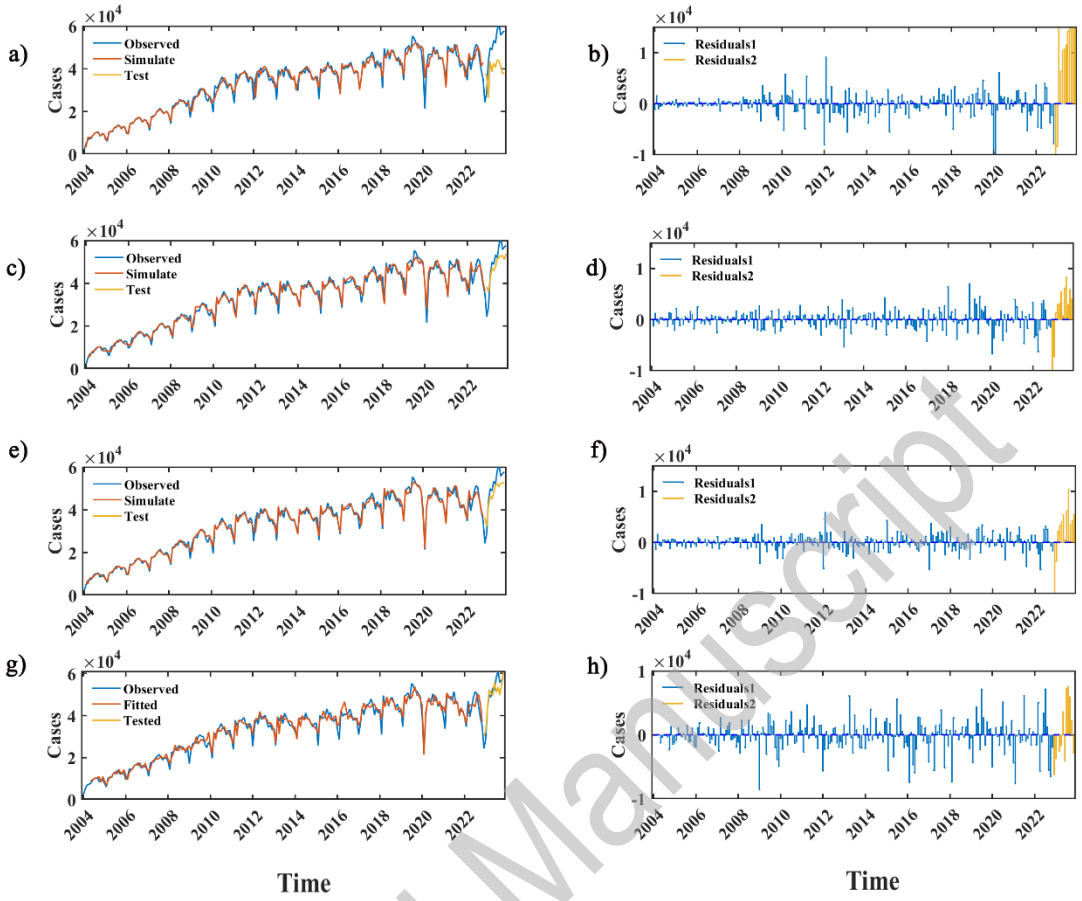


Figure 6 Fitting and forecasting performence of the SARIMA, LSTM, SARIMA-LSTM, and SARIMA-NARX models

Note: Panel a**,** c, e, and g denote the fitting and predicting results using the SARIMA, LSTM, SARIMA-LSTM, and SARIMA-NARX models, respectively, the red and yellow curves represent the simulation values for the train set and test set of the TS**.** Panel b, d, f, and h denote the residuals of the SARIMA, LSTM, SARIMA-LSTM, and SARIMA-NARX models, respectively, the blue and yellow stems represent the residuals for the train sets and test sets, respectively.

3.6 Predictions of SARIMA-LSTM and SARIMA-NARX models

We re-modeled the SARIMA, SARIMA-LSTM, and SARIMA-NARX models with all original data before predicting future time steps to ensure the accuracy of the predictions. The new SARIMA model was built first as a basis for the other two models, and after the sample size of the time series used for modeling was increased, the best fitting SARIMA model was established as SARIMA(3,1,1) $(4,1,0)_{12}$ with AIC and BIC values of 4540.9 and 4569.3, respectively. The construction of the hybrid SARIMA-LSTM and SARIMA-NARX models was then carried out based on the fitted values of the SARIMA model, and 20 times predictions of the monthly incidence of syphilis for the next 24 months (From December 2023 to November 2025) were made using the two models. In the forecasting application of the SARIMA-LSTM and SARIMA-NARX models, the architecture was preserved with an identical neuron configuration as employed during the training phase. Furthermore, an ensemble approach was implemented, where each model executed twenty times of forecasts. The extrema,

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

specifically the maximal and minimal predictive values at each temporal increment, were systematically documented. This procedure was designed to facilitate the construction of predictive interval plots, enhancing the visualization and interpretation of forecast uncertainty. The results showed that the trends predicted by the SARIMA-LSTM and SARIMA-NARX models were similar, the forecasted values of the SARIMA-NARX model are slightly higher overall than those of the SARIMA-LSTM. The peak number of monthly incidences appeared in the July and August of 2024 during the prediction period(**Figure 7**).

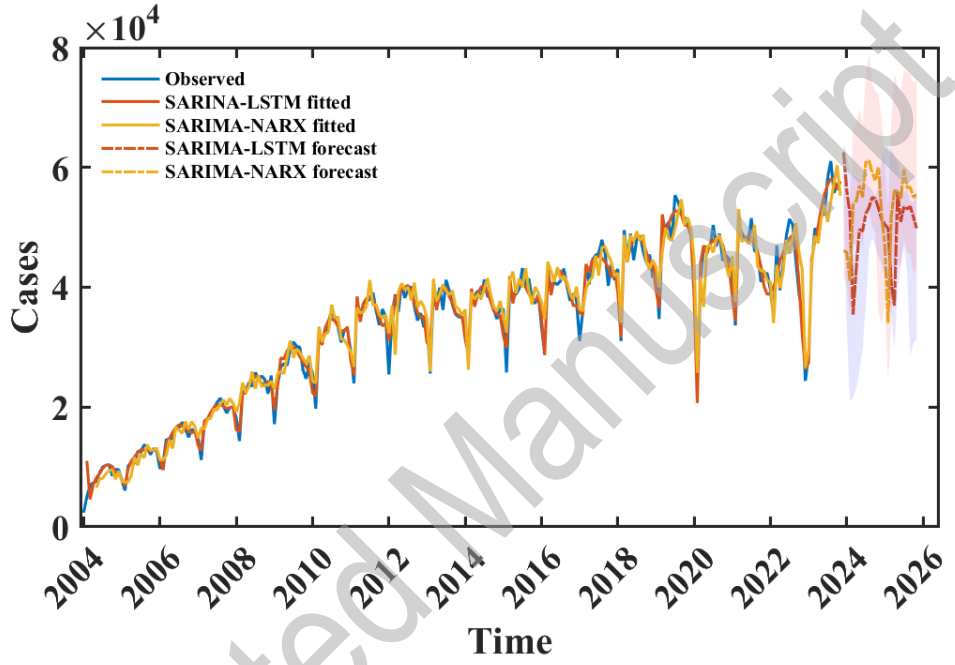


Figure 7 Prediction results from December 2023 to November 2025 of SARIMA-LSTM, and SARIMA-NARX models

Note: The light gray and blue areas respectively represent the forecast intervals of SARIMA-NARX and SARIMA-LSTM. The blue, red, and yellow curves represent the original data, the fitted values of the SARIMA-LSTM and SARIMA-NARX models, and the predicted values of the two models are represented by the red and yellow dashed lines.

4 Discussion

Syphilis is a highly insidious STD and has long-lasting damage to the human body. As the STD with the highest incidence in China and has been increasing for more than 10 years, it is essential to fit and predict the incidence data of syphilis. This will help the government to formulate relevant public health policies in advance, and rationalize the allocation of public health resources to avoid widespread infection in the population. After seasonal and long-term trend decomposition of the sample data, we found that the syphilis prevalence in China has been growing rapidly from 2004 to 2012, since then onset growth has slowed until the end of 2019. The incidence is generally higher in the summer and fall of each year, and after the outbreak of COVID-19, the growth trend slowed down but the overall incidence level is still high. Undoubtedly, the stringent

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

public health and social interventions adopted in China's response to the COVID-19 process have played a specific role in curbing the transmission of syphilis, but as China gave up the *zero-COVID* strategy for the normalization of prevention and control of the COVID-19, the epidemiology trend of syphilis need to be reassessed. In February 2020, the number of monthly reported cases of syphilis was 21,448, a decrease of 53.1% compared with January 2020, making it the month with the lowest number of reported cases in the past 10 years, probably since at this time, lockdowns and restrictions on gatherings and travel were imposed in the nationwide and most of the inhabitants and healthcare workers were amid their Lunar New Year vacations, which impacted the detection of the disease and the reporting of cases[29]. Then, China implemented the *zero-COVID* strategy, and population mobility was significantly reduced, resulting in a blockage of interpersonal transmission of the disease, which has resulted in a smooth fluctuation in the number of reported cases, rather than continuing to grow.

The seasonality of syphilis epidemic behavior can be related to sexual behaviors in Chinese populations and the patients' clinical attendance[30]. Some studies have found that in early spring, shortly after the end of the Chinese New Year, there is often a mass migration of Chinese populations, many of whom are returning from rural to urban jobs, and that this group behavior is often accompanied by an increase in sexual behavior[31,32]. Influenced by traditional Chinese beliefs, the willingness to diagnose and seek medical care around the Spring Festival is lower than at other times of the year, but in the summer months, graduated students and job seekers undergo mandatory medical check-ups before enrolling in school and the military as well as entering the workforce, which included serological testing for the virus[31], in addition, changes in hormone levels may lead to an increase in sexual behavior during specific seasons[32], leading to a peak in incidence in the summer and fall, which may explain the seasonal pattern of syphilis in China.

As a traditional mathematical model, SARIMA model is widely used in the analysis of time series data[33-35], but one of the applicable conditions is that the research data must be smooth, so it is often necessary to transform the data to achieve the modeling conditions, this process will often lose part of the information contained in the original data, for the fitting and prediction of complex nonlinear data, SARIMA model may not be able to precisely. The emergence of neural network algorithms has improved this deficiency, as the neural network model involves a large number of neurons in computation, and the overall system output is calculated through the interactions between neurons. As a result, the network possesses good robustness, such that even if there are errors in a certain part of the network, it will only reduce the adaptability of the network rather than causing significant errors.

Through iterations, the connections between neurons can be adjusted, allowing specific logical operations or nonlinear computations to be performed from complex or imprecise data. Although the LSTM and NARX models have unique advantages in time series data modeling, a single neural network model still has limitations in its usage. LSTM and NARX neural networks simply use known inputs to estimate the current

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

output, which may affect the accuracy of their predictions and inferences[10], especially when time variables are crucial, particularly in cases where time series exhibit seasonality. Therefore, we attempt to establish SARIMA-LSTM and SARIMA-NARX combined models to explain the relationship between the fitted values of the SARIMA model and the sample data, emphasizing the time variable. As the SARIMA model has a good capturing ability for periodic fluctuations, while LSTM and NARX excel in capturing nonlinear oscillations. Our research results also demonstrate that the combined model has better predictive capabilities than a single model, indicating that the combined model can integrate the strengths of individual single models.

In terms of predictive power, predictive models are considered perfect when the MAPE value is less than 5%. Models with MAPE values in the range of 5%-10% are considered high-precision models; models with MAPE values in the range of 10%-20% are considered good models[36]. Although the thoughts of building the hybrid SARIMA-LSTM and SARIMA-NARX models in this study are different, they are both a quantitative description of the relationship between the output of SARIMA and the actual onset data using the LSTM and NARX models, which can be regarded as the SARIMA model nested in the essentially neural network models and thus the results are comparable. However, in determining the best-fitting SARIMA-LSTM and SARIMA-NARX models, we found that the prediction performance of SARIMA-NARX models is not as stable as that of SARIMA-LSTM models when different parameters are used for the model construction, and therefore, there is a higher demand for parameter selection, otherwise, the prediction error will be large. For the actual data in July and August 2023, both the SARIMA-LSTM and SARIMA-NARX models exhibit underestimation, which may be attributed to the fact that the data for these two months exceeds all the data within our study period and can be considered as outliers. The predictions of the SARIMA-LSTM and SARIMA-NARX showed that the results of the two models had similar trends, suggesting that syphilis epidemiological trends in China will remain characterized by a high and stable level of epidemiological trends in the future. From May 2023 to November 2023, the reported number of cases for each month was significantly higher than the number of cases during the corresponding period, indicating the need for early intervention measures to prevent potential risks. Since the social and economic impacts of COVID-19 have not yet been eliminated, this study can provide a cost-effective tool for China and worldwide, which can help to identify trends in disease prevalence, rationalize the allocation of public health resources, avoid the waste of medical resources, and protect people's health.

5 Limitations

Admittedly, this study has several limitations. First, although the sample data were acquired from the official health administration in China, they were reported and aggregated by regional healthcare institutions at all levels, and between December 2019 and December 2022, the Chinese government has taken strict public health measures in response to the COVID-19 pandemic, which could lead to a decrease in the accessibility of people at high risk of syphilis infection to seek medical inspection, so the data may be subject to reporting bias. Second, although the LSTM model has high fitting

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

accuracy, the training progress and parameter optimization of the model require a lot of time because of the complex structure of the LSTM model. Third, the time series model can only be used for short-term prediction, and the accuracy will be reduced if a long-term prediction is performed, so the data needs to be updated frequently to optimize the model. Finally, the determination of model parameters for LSTM and NARX models currently lacks a well-established theoretical framework and often relies on heuristic and empirical methods. While model selection is typically based on the evaluation of prediction performance using a test set, this approach may not adequately assess the model's generalization ability to unforeseen data. Therefore, caution is advised when extrapolating the predictions of future disease incidence from this study.

6 Conclusions

The hybrid SARIMA-NARX and SARIMA-LSTM methods predict syphilis cases more accurately than the basic SARIMA and LSTM methods, so that can be used for governments to rationally allocate health resources and develop long-term syphilis prevention and control programs. In addition, the predicted cases still maintain a fairly high level of incidence, so there is an urgent need to develop more comprehensive prevention and control, and intervention strategies.

Supplementary material. The supplementary material for this article was uploaded the same as the manuscript.

Data availability statement. The sample data is available at *http://www.nhc.gov.cn/jkj/new_index.shtml* and https://www.cnki.net/.

Acknowledgements. Not applicable.

Financial support. (0417) Joint Training Program Of Weifang Medical University Public Health Crisis Management Doctors.

Competing interest. The authors declare none.

Ethical standard. Ethics approval was not required for this study as only open data were used (no human participants were involved in this study).

Informed consent. Not applicable (no human participants were involved in this study).

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

References

(1) **Kenyon C, et al.** Management of asymptomatic sexually transmitted infections in Europe: towards a differentiated, evidence-based approach. *The Lancet regional health Europe,* 2023; **34**: 100743.

(2) **Hook EW, 3rd.** Syphilis. *Lancet (London, England),* 2017; **389**(10078): 1550-1557.

(3) **WHO.** Overview of syphlis. *https://wwwwhoint/health-topics/syphilis#tab=tab 1,* 2023; (accessed 9 August 2023).

(4) **Mercuri SR, et al.** Syphilis: a mini review of the history, epidemiology and focus on microbiota. *The new microbiologica*, 2022; **45**(1): 28-34.

(5) **Chauhan K, et al.** Demystifying Ocular Syphilis-A Major Review. *Ocular Immunology and Inflammation,* 2023; **31**(7): 1425-1439.

(6) **Sankaran D, et al.** Congenital Syphilis-An Illustrative Review. *Children (Basel),* 2023; **10**(8):1310.

(7) **Wozniak PS, et al.** The Mortality of Congenital Syphilis. *Journal of Pediatrics,* 2023; **263**:113650.

(8) **Zhang M,et al.** Factors associated with adverse pregnancy outcomes of maternal syphilis in Henan, China, 2016–2022. *Epidemiology and Infection,* 2023; **151**:e170.

(9) **National Disease control and prevention Administration,** 2021 National Overview of Statutory Infectious Disease Epidemics, https://www.ndcpa.gov.cn/jbkzzx/c100031/common/content/content_1651466519851110400.html, 2022; (accessed 22 February 2024).

(10) **Huang J, et al.** Spatial-temporal analysis of HIV/AIDS and syphilis in mainland China from 2007 to 2017. *Journal of medical virology,* 2022; **94**(7): 3328-3337.

(11) **Wang KW, et al.** Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network. *Epidemiology and infection,* 2017; **145**(6): 1118-1129.

(12) **Wang M, et al.** ARIMA and ARIMA-ERNN models for prediction of pertussis incidence in mainland China from 2004 to 2021. *BMC public health*, 2022; **22**(1): 1447.

(13) **Choi RY, et al.** Introduction to Machine Learning, Neural Networks, and Deep Learning. *Translational vision science & technology,* 2020; **9**(2): 14.

(14) **Kriegeskorte N, Golan T.** Neural network models and deep learning. *Current biology : CB,* 2019; **29**(7): R231-r236.

(15) **Yu Y, et al.** A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural computation,* 2019; **31**(7): 1235-1270.

(16) **Dhafer AH, et al.** Empirical Analysis for Stock Price Prediction Using NARX Model with Exogenous Technical Indicators. *Computational intelligence and neuroscience*, 2022; **2022**: 9208640.

(17) **Zhang WR, et al.** Forecasting groundwater level of karst aquifer in a large mining area using partial mutual information and NARX hybrid model. *Environmental research,* 2022; **213**: 113747.

(18) **Wang Y, et al.** Seasonality and trend prediction of scarlet fever incidence in mainland China from 2004 to 2018 using a hybrid SARIMA-NARX model. *PeerJ,* 2019; **7**:

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

e6165.
(19) **Wang Y, et al.** Temporal trends analysis of human brucellosis incidence in mainland China from 2004 to 2018. *Scientific reports,* 2018; **8**(1): 15901.
(20) **Wang Y, et al.** Time series modeling of pertussis incidence in China from 2004 to 2018 with a novel wavelet based SARIMA-NAR hybrid model. *PloS one,* 2018; **13**(12): e0208404.
(21) **Xu D, et al.** Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environmental science and pollution research international,* 2022; **29**(3): 4128-4144.
(22) **Wen T, et al.** Modeling and forecasting CO(2) emissions in China and its regions using a novel ARIMA-LSTM model. *Heliyon,* 2023; **9**(11): e21241.
(23) **Zhao R, et al.** A hybrid model for tuberculosis forecasting based on empirical mode decomposition in China. *BMC infectious disease,* 2023; **23**(1): 665.
(24) **Wang L, et al.** Emergence and control of infectious diseases in China. *Lancet (London, England),* 2008; **372**(9649): 1598-1605.
(25) **Dhafer AH, et al.** Empirical Analysis for Stock Price Prediction Using NARX Model with Exogenous Technical Indicators. *Computational intelligence and neuroscience,* 2022; **2022**: 9208640.
(26) **Wang Y, et al.** Development and evaluation of a deep learning approach for modeling seasonality and trends in hand-foot-mouth disease incidence in mainland China. *Scientific reports*, 2019; **9**(1): 8046.
(27) **Smith LN.** Cyclical Learning Rates for Training Neural Networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV),* 2017; 464-472.
(28) **Wang Y, et al.** Temporal trends analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a new SARIMA-NARNNX hybrid model. *BMJ open,* 2019; **9**(7): e024409.
(29) **Wu YL, et al.** Impact of COVID-19 epidemic on syphilis case reporting in China. *Chinese Journal of Epidemiology,* 2022; **43**(12): 2015-2020.
(30) **Wright RA, Judson FN.** Relative and seasonal incidences of the sexually transmitted diseases. A two-year statistical review. *The British journal of venereal diseases,* 1978; **54**(6): 433-440.
(31) **Li X, et al.** HIV/AIDS-related sexual risk behaviors among rural residents in China: potential role of rural-to-urban migration. *AIDS education and prevention : official publication of the International Society for AIDS Education,* 2007; **19**(5): 396-407.
(32) **Zhang X, et al.** Study of surveillance data for class B notifiable disease in China from 2005 to 2014. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases,* 2016; **48**: 7-13.
(33) **Tan NX, et al.** Temporal trends in syphilis and gonorrhea incidences in guangdong province, china. *The Journal of infectious diseases,* 2014; **209**(3): 426-430.
(34) **Duangchaemkarn K, et al.** SARIMA Model Forecasting Performance of the COVID-19 Daily Statistics in Thailand during the Omicron Variant Epidemic. *Healthcare (Basel, Switzerland),* 2022; **10**(7).

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press

(35) **Wang S, et al.** Comparison of SARIMA model and Holt-Winters model in predicting the incidence of Sjögren's syndrome. *International journal of rheumatic diseases,* 2022; **25**(11): 1263-1269.
(36) **Tan CV, et al.** Forecasting COVID-19 Case Trends Using SARIMA Models during the Third Wave of COVID-19 in Malaysia. *International journal of environmental research and public health,* 2022; **19**(3).
(37) **Pao H-T.** Forecasting energy consumption in Taiwan using hybrid nonlinear models. *Energy,* 2009; **34**(10): 1438-1446.

https://doi.org/10.1017/S0950268824000694 Published online by Cambridge University Press