HIV/AIDS - Research and Palliative Care Dovepress

Open Access Full Text Article ORIGINAL RESEARCH

# Study on Univariate Modeling and Prediction Methods Using Monthly HIV Incidence and Mortality Cases in China

Yuxiao Yang[1,2], Xingyuan Gao[3], Hongmei Liang[4], Qiuying Yang[1,2]

[1]School of Biomedical Engineering, Capital Medical University, Beijing, People's Republic of China; [2]Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, Capital Medical University, Beijing, People's Republic of China; [3]Design Department, Beijing HANHAIZHONGJIA Hydraulic Machinery Co., Ltd, Beijing, People's Republic of China; [4]Nursing Department, China Railway 17th Bureau Group Central Hospital, Taiyuan, People's Republic of China

Correspondence: Qiuying Yang, School of Biomedical Engineering, Capital Medical University, Beijing, People's Republic of China, Tel +86 13691283439, Email yangqiuying@ccmu.edu.cn



**Purpose:** AIDS presents serious harms to public health worldwide. In this paper, we used five single models: ARIMA, SARIMA, Prophet, BP neural network, and LSTM method to model and predict the number of monthly AIDS incidence cases and mortality cases in China. We have also proposed the LSTM-SARIMA combination model to enhance the accuracy of the prediction. This study provides strong data support for the prevention and treatment of AIDS.

**Methods:** We collected data on monthly AIDS incidence cases and mortality cases in China from January 2010 to February 2024. Among them, for modeling, we used data from January 2010 to February 2021 and the rest for validation. Treatments were applied to the dataset based on its characteristics during modeling. All models in our study were performed using Python 3.11.6. Meanwhile, we used the constructed model to predict monthly incidence and mortality cases from March 2024 to July 2024. We then evaluated our prediction results using RMSE, MAE, MAPE, and SMAPE.

**Results:** The deep learning methods of LSTM and BPNN outperform ARIMA, SARIMA, and Prophet in predicting the number of mortality cases. When predicting the number of AIDS incidence cases, there is little difference between the two types of methods, and the LSTM method performs slightly better than the rest of the methods. Meanwhile, the average error in predicting AIDS mortality cases is significantly lower than in predicting AIDS incidence cases. The LSTM-SARIMA method outperforms other methods in predicting AIDS incidence and mortality.

**Conclusion:** Due to the different characteristics of the AIDS incidence and mortality cases series, the performance of distinct methods is slightly different. The AIDS mortality series is smoother than the incidence series. The combined LSTM-SARIMA model outperforms the traditional method in prediction and the LSTM method alone, which is of practical significance for optimizing the prediction results of AIDS.

**Keywords:** AIDS, ARIMA model, Prophet model, deep learning model, LSTM-SARIMA combination model

## Introduction

AIDS has become a worldwide severe public health and social problem for 40 years since its discovery in 1981.[1] The first case of AIDS was found in China in 1985.[2] According to Global HIV & AIDS statistics—FACT SHEET issued by the Joint United Nations Programme on HIV/AIDS (UNAIDS) on World AIDS Day 2023, the number of people living with HIV went up to about 39 million [33.1 million to 45.7 million] in 2022; About 1.3 million [1 million to 1.7 million] people were newly infected in 2022 and about 630,000 [480,000 to 880,000] died from AIDS-related diseases.[3] As of June 30, 2023, there were about 1260.9 thousand living infected people and 437.3 thousand mortality cases in China.[4] AIDS seriously threatens human life health and public health, and the Chinese Government has introduced a variety of preventive and control measures. In 1995, China established the National Notifiable Disease Report System (NNDRS) and used the system to monitor the incidence of AIDS in 2004.[5]

Received: 2 May 2024
Accepted: 5 October 2024
Published: 24 October 2024

An essential strategy in HIV/AIDS surveillance, prevention and control involves the early identification of abnormal prevalence and growth trends, as well as the prompt detection and treatment of affected individuals.[6] Consensus on diagnosis and management of immunological non-responder in acquired immunodeficiency syndrome (version 2023)[7] points out that early initiation of antiretroviral treatment for AIDS patients is an essential means of improving the AIDS treatment effectiveness. By analyzing the trend of HIV infection and transmission, we can create a more accurate prediction model for AIDS, which can help in detecting HIV early, guiding clinicians in making informed decisions about diagnosis, treatment, and management, and providing a rationale theoretical basis for formulating preventive measures and allocating medical resources. So, modeling and predicting monthly incidence and mortality cases has significant practical implications. In the current study of AIDS incidence, researchers have realized a variety of single and combined prediction methods.

In the current single-method research, both traditional modeling methods and deep learning methods can achieve some results. Xu Bin et al made the ARIMA model to predict the incidence and mortality of AIDS in China.[8] Luo Zixiao et al used SARIMA and Prophet models to study the incidence of AIDS in Henan Province.[9] Zhao Tianming et al then used the SARIMA model to discuss the data impact of COVID-19 policy on AIDS.[10] Wu Hailai et al and Li et al demonstrated that BP-ANN performs well in the analysis and prediction of AIDS,[11,12] respectively. Wang G et al found that LSTM models generally have higher prediction accuracy than traditional models.[13]

Similar to most infectious diseases, the spread of AIDS has both a considerable degree of nonlinear characteristics and linear characteristics.[14] To reasonably utilize the advantages of each model, some scholars chose to use a combination of multiple modeling methods, which also provides new ideas and development directions for AIDS prediction. For example, Yawen Wang et al established an autoregressive moving average model and a combination model of ARIMA and GRNN (generalized regression neural network).[15] An Qingyu et al proposed an EMD (empirical modal decomposition)-BPNN combined model.[16] Zixiao Luo et al used a SARIMA-Prophet combined model based on the L1 paradigm to study the incidence of AIDS in Henan Province.[9] Ying Chen et al combined LSTM-ARIMA to predict the incidence of AIDS among children in East Asia.[17] However, these models also have certain limitations, such as limited data coverage, fewer comparable models, and an inability to study multifarious data with different characteristics. Therefore, the existing models generally suffer from low prediction accuracy and poor generalization ability.

Based on the above issues, in order to establish a more accurate and applicable AIDS prediction model, assist relevant medical workers in understanding the development trend of the AIDS epidemic in advance, and reduce the economic burden of AIDS on individuals and the community, this study uses the monthly AIDS incidence cases and mortality cases in China for a total of more than 14 years and 170 months from January 2010 to February 2024 as the research sequences. We give a modeling and forecasting study combining the decomposition of sequences into trend, seasonal, and trending sequences under the advance of sufficiently analyzing the two types of data's self-characteristics, such as volatility, cyclicity, seasonality, trend, etc., and at the same time, combining the advantages of a single model, ARIMA, SARIMA, Prophet, BP neural network, LSTM, and so on, in terms of their respective capabilities in processing data. Meanwhile, considering the overfitting problem of residual terms caused by the high tracking nature and high sensitivity of deep learning models to the trend of sequences, combining the processing ability of SARIMA on residual sequences, the study researchers explored the LSTM-SARIMA combination model, and using the different normalization methods according to the characteristics of the sequences. The LSTM-SARIMA combination model achieves better forecast accuracy.

## Methods

In this study, we use ARIMA, SARIMA, Prophet, BP Neural Network, LSTM five-single methods, and one combination method LSTM-SARIMA to achieve the prediction. Using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (SMAPE) to evaluate the results.

In LSTM-SARIMA, the researcher participants use Min-Max normalization to normalize the original time series and the decomposed trend series and use Sigmoid to normalize the residual series from the data decomposition.

## Data Description

Using a total of 170 months of data from the National Disease Control and Prevention Administration in China, we included Monthly Incidence and Mortality cases data from January 2010 to February 2024 for more than 14 years,[18,19] which is shown in Figure 1.

## ARIMA Model and SARIMA Model

ARIMA (Autoregressive Integrated Moving Average Model), namely the Box–Jenkins model, is a classic and popular time-series forecasting model proposed by the American statisticians Jenkins and Box, also known as the Box–Jenkins method.[20] The basic idea of the model is to use the historical information of the series itself to forecast the future. It first extracts the time change law of the sequence through the autocorrelation and difference of the sequence and then predicts the future trend of the sequence according to the change law.

ARIMA model can be generally denoted as ARIMA ($p$, $d$, $q$), which has three basic parts: autoregressive (AR), moving average (MA), and single integer order I. ARIMA is expressed as:[21]

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^d Y_i = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_i \tag{1}$$

In the formula, $L$ means the lag operator, $p$ means the autoregressive coefficient, $q$ means the moving average coefficient, and $d$ is the number of differences needed to smooth the time series.

The implementation of the ARIMA model is as follows:

① Stabilization test: Using the Augmented Dickey–Fuller (ADF) test to examine the stabilization of the original sequence,[22] make a difference if the sequence is unstable, and repeat the difference to stabilize the sequence. The number of times of difference is the value of parameter $d$.

② Parameters estimation of $p$, $q$: using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) tests on the stabilization series to determine the range value of $p$, $q$, applying the Akaike Information Criterion (AIC)[23] and Bayesian Information Criterion (BIC) on the parameters within the range, and take the parameter results that are smaller in both as alternative parameters.

③Model check: Alternative parameters were entered into the model, respectively, to test the significance of the parametric model estimates and to comparatively check the model's predicted data for apparent errors to determine the final model parameters.[24]

④ Model forecast: using the parameters identified in ③ to forecast the subsequent values of the sequence.

ARIMA model parameters are adjustable. In the sequence calculation, we took the historical trend of this model into account, and the forecasting accuracy is high. However, there are certain deficiencies in handling periodic series.

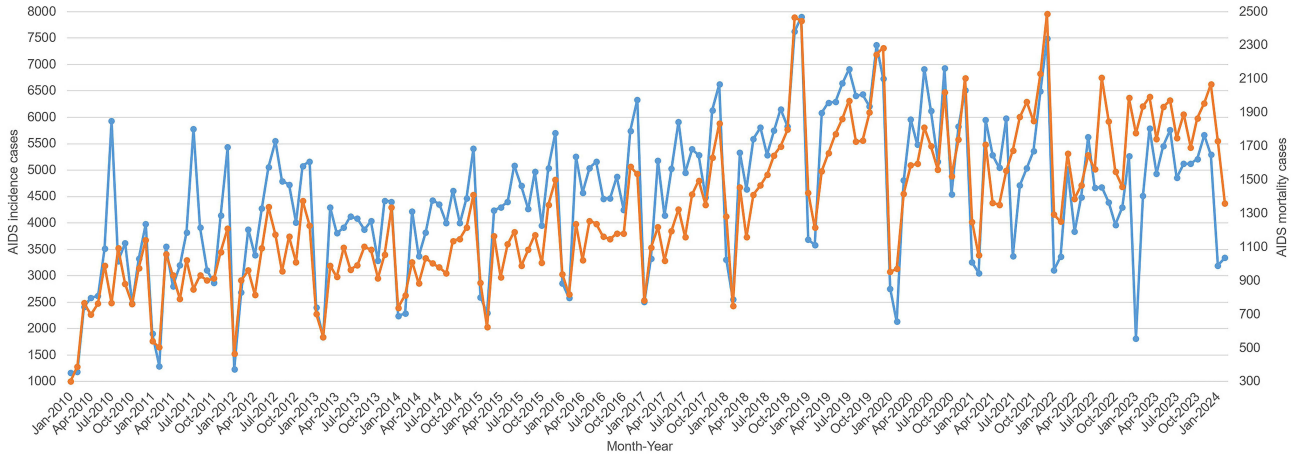


**Figure 1** AIDS monthly incidence cases and mortality cases in China, Jan-2010–Feb-2024. (Blue line = AIDS incidence cases per month, Orange line = AIDS mortality cases per month).

Therefore, we extracted the periodicity from the original sequence and decomposed it into three sequences: trend, periodic, and residual. Then, we used the ARIMA method on the trend and residual sequences separately and summed their predictions with the period sequence. Finally, we get the forecast data of the ARIMA model.

Compared to ARIMA, Seasonal Autoregressive Integrated Moving Average (SARIMA) adds seasonal parameters and improves the ability to handle seasonal component time series. It is better than the universal ARIMA in long-term and periodic time series forecasting.

The SARIMA model can be generally expressed as SARIMA (*p, d, q*) (*P, D, Q*) *s*, whose expression is formula 2:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)\left(1-\sum_{i=1}^{p}A_i L_s^i\right)(1-L)^d(1-L_s)^D Y_i = \left(1+\sum_{i=1}^{q}\theta_i L^I\right)\left(1+\sum_{i=1}^{Q}B_i L_s^i\right)\in_i \quad (2)$$

In the formula, *P, D*, and *Q* are seasonal hyperparameters, and $L_s$ is the seasonal cycle.

The implementation of the SARIMA model is similar to ARIMA, that is, using the difference to obtain the *d* and *D*, grid search method, or auto-ARIMA to get *p, q, P*, and *Q*.

## Prophet Model

Prophet is a time-series forecasting tool that combines machine learning fitting and time-series decomposition based on time and variable values. The model adds the fitting of noise terms and holidays while taking into account the effects of trend, seasonality, and periodicity, and its expression is formula (3):[25]

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \quad (3)$$

In the formula, *t* is the moment, y(t) is the time series to be decomposed, g(t) is a trend term to characterize the non-periodic trend of the time series, s(t) is a seasonal term to feature the impact with fixed periodicity, h(t) is a holiday term to characterize the effect of the time series on a specific date, and $\varepsilon(t)$ is an error term obeying a Gaussian distribution.

In the Prophet model, there are two widely used trend term g(t) functions. One is based on the logistic regression function, and the other is based on the segmented linear function.[26]

Logistic regression function is formula (4):

$$f(x) = \frac{c}{1+e^{-k(x+m)}} \quad (4)$$

Segmented linear functions is formula (5):

$$f(x) = kt + m \quad (5)$$

In the formula (4), *C* means the carrying capacity. In the formulas (4) and (5), *k* is the growth rate, and *m* is the offset parameter.

In the study, we employed the logistic regression function as the trend term function in the prophet model.

## Neural Networks Model

Alexander Bain (1873) and William James (1890) proposed the fundamentals of neural networks. As a machine learning algorithm, it can simulate the function of the human brain's nervous system. Moreover, using the connection and computation of multiple nodes (ie, neurons) can realize the combination and output of nonlinear models. Its structure mainly includes the input layer, output layer, hidden layer, optimizer, and so on. Neural networks can be categorized into two types according to the network structure: feed-forward neural networks and feedback neural networks. The signal of feed-forward neural networks generally can only propagate along the direction of the output, and the output of each moment is only related to the input of the current moment; the signal of feedback neural networks has some loops in the transmission, and the output results between the moments will be not only determined by the input of the current moment but also will be influenced by each other through the feedback loops.

Feedforward neural networks mainly include BP neural networks (Back Propagation Neural Networks, BPNN), Deep Neural Networks (Deep Neural Networks, DNN), Convolutional Neural Networks (Convolutional Neural Networks, CNN), and so on. Feedback Neural Networks mainly include Recurrent Neural Networks (RNN) and Hopfield neural networks.

In this study, we use the BP neural network of the feedforward neural network and Long Short-Term Memory (LSTM) in the RNN of the feedback neural network to analyze the data.

## BPNN Model

Rumelhart and McClelland proposed the BP Neural Network in 1986. As a multilayer feedforward neural network, the BPNN is trained according to the error back-propagation algorithm.[27] The topology of the BPNN with error back-propagation is shown in Figure 2. This network is one of the most widely used neural network models. BP network uses a back-propagation algorithm to determine the network node weights compared to traditional feedforward neural networks. In this study, the connection weights and thresholds of the BP network are first randomly initialized, the thresholds and connection weights are adjusted using the training error of the actual data, and the final neural network parameters are determined by gradient descent and minimizing the training error.[28]

The process of the BPNN includes three steps: hidden layer selection, forward transfer sub-process construction, and reverse transfer sub-process construction.

1. The number of hidden layer nodes selected uses the empirical formula (6):

$$h = \sqrt{m+n} + a \tag{6}$$

1. In the formula, $h$ is the number of nodes in the hidden layer, $m$ is the number of nodes in the input layer, $n$ is the number of nodes in the output layer, and $a$ is a value between 1 and 10 conditioning constant.
2. Forward transfer sub-process construction uses an activation function. There is no threshold for the input layer nodes in BPNN.
3. The weights and thresholds between the implicit and output layers, the input layer, and the implicit layer of the reverse transfer sub-process are calculated as follows:

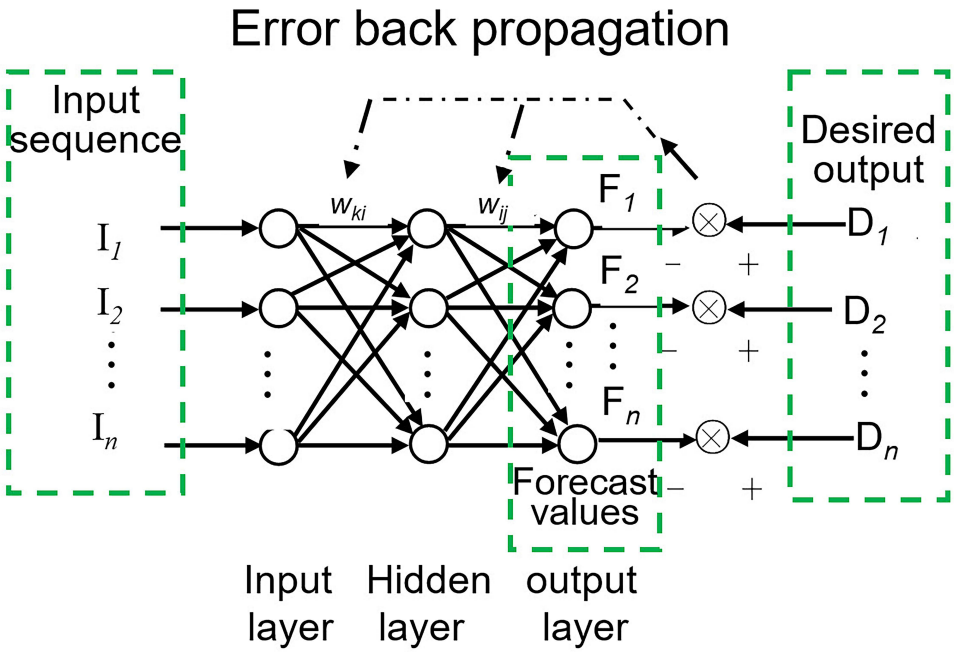


**Figure 2** The topology of the BPNN with error back-propagation.

$$w_{ij} = w_{ij} - \eta_1 * \delta_{ij} * x_i \tag{7}$$

$$b_j = b_j - \eta_2 * \delta_{ij} \tag{8}$$

$$w_{ki} = w_{ki} - \eta_1 * \delta_{ki} * x_k \tag{9}$$

$$b_i = b_i - \eta_2 * \delta_{ki} \tag{10}$$

In the formula (7) to (10), $i$ is the hidden layer node, $j$ is the output layer node, $k$ is the input layer node, $w$ is the weight between nodes, $b$ is the threshold, $x$ is the node output, $\delta$ is the learning rule between nodes, and $\eta$ is the proportion of weight vector correction.

In practice, BPNN has the characteristics of parallelism, nonlinearity, and perturbation resistance, so it has a faster learning speed, greater fault tolerance, and higher prediction accuracy compared to traditional methods; due to the higher complexity of the RNN model, the phenomena of gradient vanishing and gradient explosion are prone to occur when the input variable enters into the implicit layer along the direction of signal transmission.[29]

## LSTM Model

Long Short-Term Memory Network (LSTM) is a special variant of RNN.[30] Hochreiter and Schmidhuber first proposed the model, which corrects the disadvantage of RNN that produces gradient explosion and gradient vanishing[31] and improves the problem that ordinary RNN neural networks tend to forget long-term sequential information by constructing memory storage units. LSTM consists of five main components: cell state, hidden state, input gate, output gate, and forgetting gate.

- Cell state: the internal state of an LSTM memory cell;
- Hidden state: an internal hidden state that is devoted to computing external forecasts;
- Input gate: determines the input value that will be fed into the cell state;
- Output gate: decides the number of cell states that will be output from the hidden state;
- Forget gate: determines the previous cell state and the quantity that will be fed into the current cell state.

The first run in the process of the LSTM model is the forget gate. The expression is the formula:

$$F_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{11}$$

Then, the information will pass the input gate to determine the amount of input that should be fed into the cell state in the current situation. The expression is the formula:

$$I_t = \sigma\left(W_i \times [h_{t-1}, x_t] + b_f\right) \tag{12}$$

$$\hat{C} = \tanh\left(W_c \times \left[h_{t-1,} x_t\right] + b_c\right) \tag{13}$$

Finally, the output gate gives the output value and synthesizes the inputs and outputs in the past time and the current state of the unit to obtain the final forecasted value. The expression is the formula:

$$C_t = f_t \times C_{t-1} + i_t \times \hat{C}_t \tag{14}$$

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{15}$$

$$h_t = O_t \times tanh(C_t) \tag{16}$$

In the formula (11)-(16), σ represents the activation function, $\hat{C}_t$ represents the candidate vector, $W$ represents the weight, and $b$ represents the bias; $f,c,i,o$ represent the forget gate, the candidate vector, the input gate, and the output gate, respectively; $h_t$ is the output (the state of the hidden layer) at the current moment, $C_t$ is the state at this moment in time, and $O_t$ denotes the portion of the output gate output, $I_t$ denotes the portion of the input gate output, $F_t$ denotes the portion of the forget gate output.

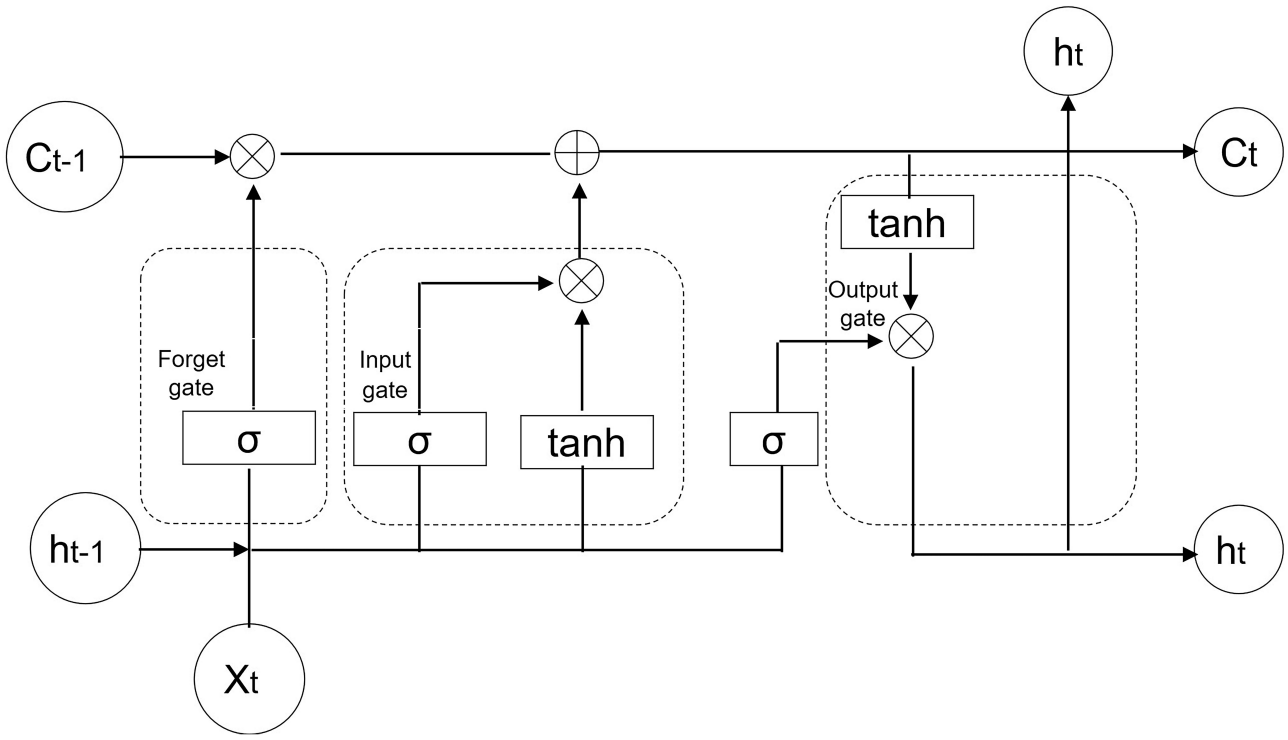


**Figure 3** Neuron structure of LSTM model.

The neuron structure of the LSTM model is shown in Figure 3.

The basic structure of LSTM is more complex than the standard RNN. LSTM has two different states, the cell state and the hidden state,[32] while the traditional RNN has only one hidden state. At the same time, LSTM has three different gates to better process the previous cell states and current outputs to compute more reasonable hidden state parameters. Thus, it can meet the demand for higher forecast accuracy.

## Normalization and Evaluation Criteria

### Normalization

The basic structure of LSTM is more complex than the standard RNN. LSTM has two different states, the cell state and the hidden state,[33] while the traditional RNN has only one hidden state. At the same time, LSTM has three different gates, which allow for a better processing of the previous cell state and the current output, thereby calculating a more accurate and effective hidden output state. Thus, it can meet the demand for higher forecast accuracy.

This study uses Min-Max and Sigmoid normalization methods to process the data.

Min-Max normalization, namely outlier normalization, can convert values to 0–1 by certain linear transformations, thus reducing the adverse effects caused by the outliers of the data. The formula for this normalization is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{17}$$

In the formula, $x$ is the original value, min(x) is the minimum value of the sequence, min(x) is the maximum value of the sequence, and x' is the normalized value.

Therefore, sometimes, we can replace the maximum and minimum values with given values, which avoids data recovery difficulties caused by changing the normalization parameters from sample to sample.

In our study, the residuals sequence has more distinguished values. Sigmoid normalization allows the model to pay more attention to the cyclicality and tending of the sequences to optimize the output results.

The expression of the Sigmoid normalization is formula (18):

$$x' = \frac{t}{1 + e^{x}} \tag{18}$$

In the formula, t is the deflation parameter, x is the current data, and x' is the normalized data.

### Evaluation Criteria

The study evaluated the models in two ways: comparing their prediction accuracy under the same data conditions and assessing their applicability using prediction results from different types of data. Therefore, the study participants use RMSE, MAE, MAPE, and SMAPE as model training evaluation criteria.

Mean Absolute Error (MAE) calculates the mean of the error between the predicted output and the actual value. The smaller the MAE, the better the forecast.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y_i| \tag{19}$$

Root Mean Square Error (RMSE) is the square of the mean square error. This value incorporates a squaring operation when calculating the error, so larger errors have a more significant effect than the MAE. The RMSE provides a more intuitive reality of the dispersion of the forecast error while reflecting the accuracy of the forecast.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{20}$$

Mean Absolute Percentage Error (MAPE) is calculated by adding the actual data as the denominator and is a better visualization of the accuracy of the forecast.

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{21}$$

Symmetric Mean Absolute Percentage Error (SMAPE). The smaller the value, the better the fit of the forecast to the actual value.

$$\mathrm{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{(|\hat{y}_i| + |y_i|)/2} \tag{22}$$

In the formulas (19)-(22), n is the total amount of data, $y_i$ is the actual data, and $\hat{y}_l$ is the forecasted data.

The joint use of MAE and RMSE metrics is often employed to examine the dispersion of sample error. By comparing the difference between these two values across various models, we can clarify the predictive accuracy of the models and reduce the impact of extreme values on model assessment. Furthermore, MAPE and SMAPE can be used with MAE to estimate the degree of fit for samples of different orders of magnitude. In this study, specifically, we utilized MAE, RMSE, MAPE, and SMAPE to assess the differences in predictions between the number of AIDS incidence and mortality cases, as well as to clarify the influence of data distribution characteristics on the predictive validity of our model.

All models in our study are performed using Python 3.11.6.

## Results

This study uses 134 months of data from January 2010 to February 2021 in the dataset as the training set and 36 data from March 2021 to February 2024 as the test set. Meanwhile, we use the constructed model to predict the number of monthly incidence cases and mortality cases for five months from March 2024 to July 2024.

Figure 4 shows the flow chart of models. Figure 4A shows the traditional modeling method flow chart. Among them, the black dashed line part is the Standard ARIMA method, the green line part is the ARIMA method after sequence decomposition, the red line part is the Prophet method, and the orange line part is the SARIMA method. Figure 4B shows

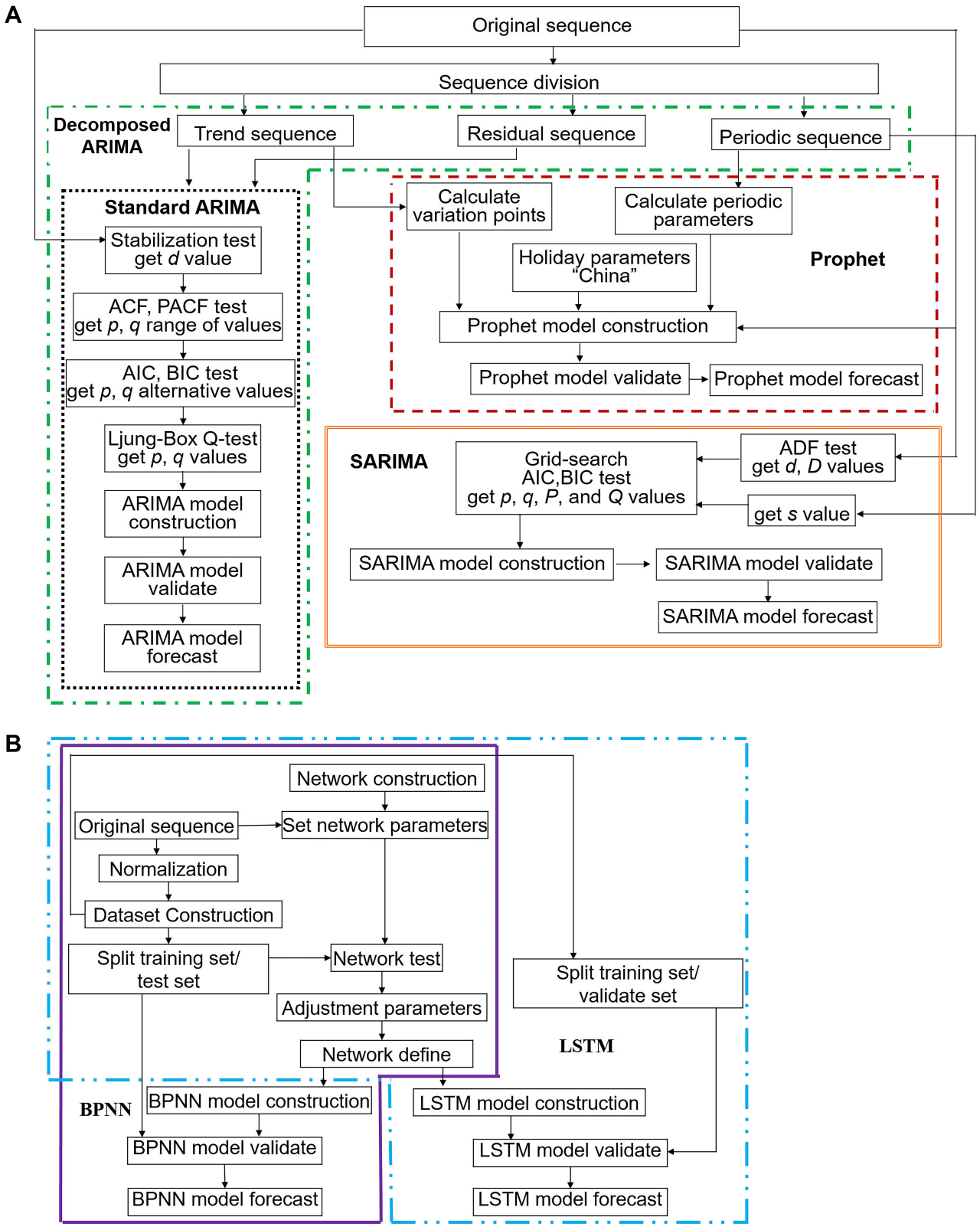


**Figure 4** Continued.

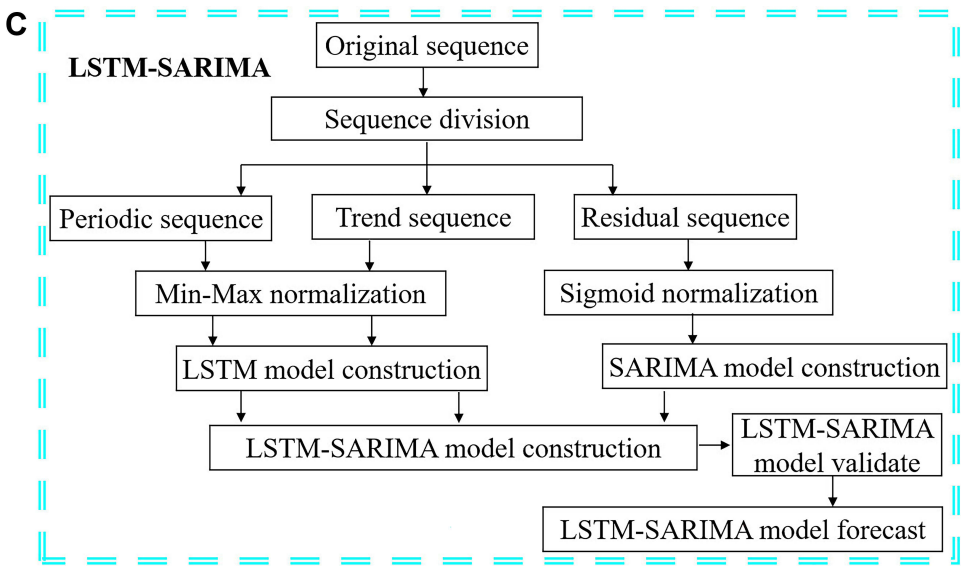


**Figure 4** Flow chart of models. (**A**) Traditional modeling method. (Black dashed line part=Standard ARIMA, Green line part= Decomposed ARIMA, Red line part=Prophet, Orange line part=SARIMA). (**B**) Deep learning method. (Purple line part=BPNN, Blue line part=LSTM). (**C**) LSTM-SARIMA combined model. (Blue green line part=LSTM-SARIMA).

the deep learning method flow chart. Among them, the purple line section is the BPNN method, and the blue line section is the LSTM method. Figure 4C shows the LSTM-SARIMA combined model.

The forecast results of each model for AIDS monthly incidence cases and mortality cases are described by evaluation criteria shown in Table 1. Among them, Table 1a is the result of incidence cases, and Table 1b is the result of mortality cases. Among others, Decomposed ARIMA model training uses the sequence after the decomposition processing of the original sequence; the LSTM-SARIMA model uses the LSTM method to analyze the trend sequence, and meanwhile, the SARIMA method to analyze cycle residual sequence after the decomposition of the original sequence.

**Table 1** Evaluation Criteria Values for All Model Forecast

| Model / Evaluation criteria | Decomposed ARIMA | SARIMA | Prophet | BPNN | LSTM | LSTM-SARIMA |
|---|---|---|---|---|---|---|
| **a AIDS monthly incidence cases** | | | | | | |
| **MAE** | 690.41 | 953.42 | 707.06 | 704.89 | 662.66 | 507.87 |
| **RMSE** | 834.96 | 1161.44 | 933.32 | 904.97 | 834.16 | 682.80 |
| **MAPE** | 15.67% | 22.02% | 16.91% | 17.65% | 15.74% | 11.63% |
| **SMAPE** | 15.75% | 19.07% | 14.94% | 15.27% | 14.31% | 11.03% |
| **b AIDS monthly mortality cases** | | | | | | |
| **MAE** | 332.18 | 198.10 | 257.25 | 188.78 | 185.46 | 167.62 |
| **RMSE** | 394.51 | 251.60 | 307.93 | 229.02 | 224.89 | 205.27 |
| **MAPE** | 18.32% | 11.38% | 15.78% | 11.21% | 11.19% | 9.94% |
| **SMAPE** | 20.86% | 11.41% | 14.28% | 10.84% | 10.72% | 9.76% |

**Abbreviations**: MAE, Mean Absolute Error; RMSE, Root Mean Square Error; MAPE, Mean Absolute Percentage Error; SMAPE, Symmetric Mean Absolute Percentage Error; LSTM-SARIMA, Combined model of LSTM and SARIMA.

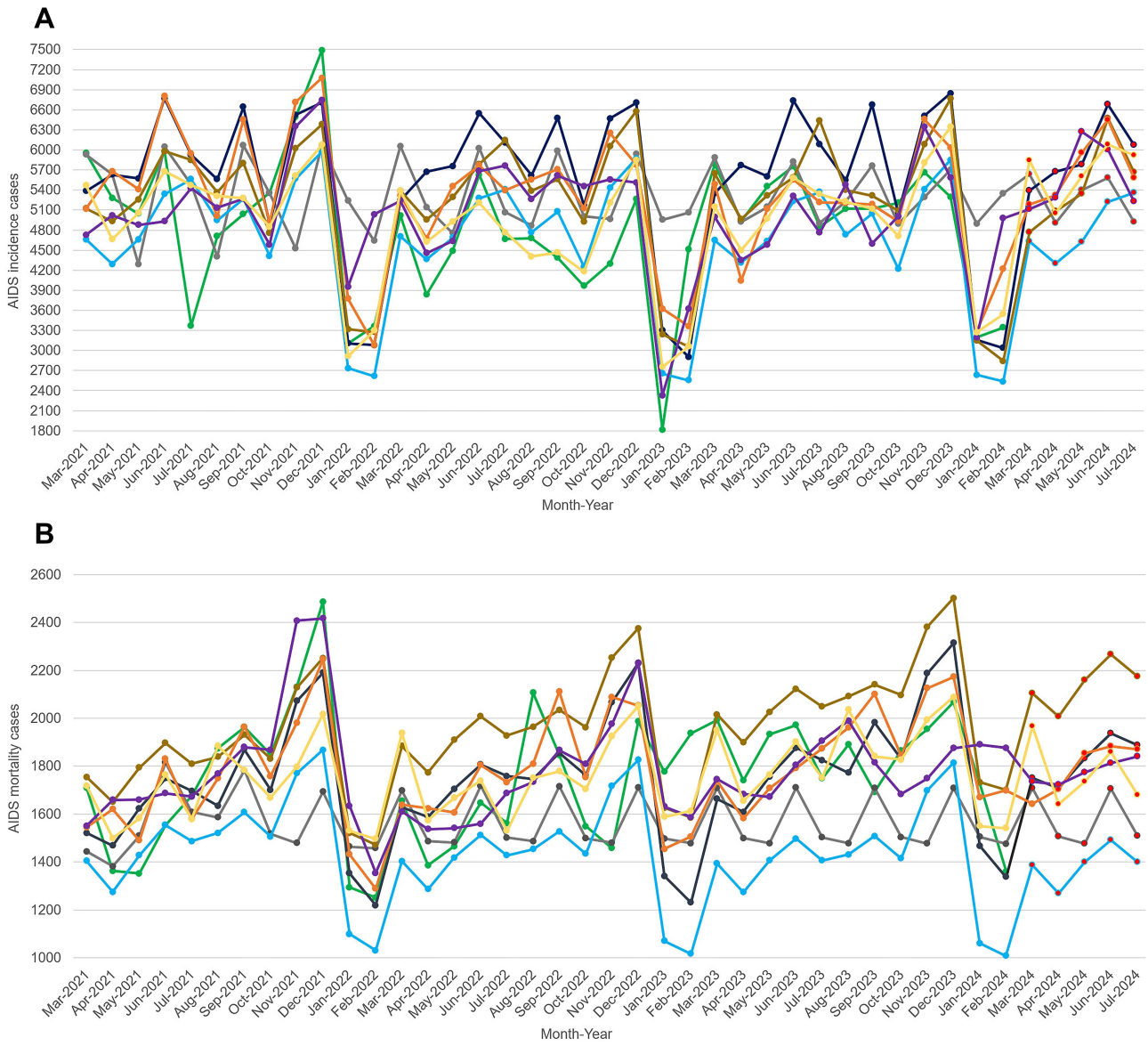


**Figure 5** Model validation and forecast results (Green line= Original sequence, Gray line = ARIMA, Blue line = Decomposed ARIMA, Black line = SARIMA, Brown line = Prophet, Orange line = BPNN, Purple line = LSTM, Gold line = LSTM-SARIMA, Red dots= Predicted values of monthly incidence and mortality cases from March to July 2024 for each model). (**A**) Incidence cases results. (**B**) Mortality cases results.

Figure 5 displays the validation and forecast results of all the models. Figure 5A shows the results of incidence cases, while Figure 5B displays the results of mortality cases.

Realization for all models is as follows:

## ARIMA Model

Using ADF to test the original sequence and the sequence of each-order difference in succession, we obtained that the t-value of the AIDS incidence cases and AIDS mortality cases are little than each critical value at the first-order difference, the significance level $p < 0.05$, and the sequences are smooth, so, $d = 1$. Using ACF, PACF test on the first-order difference, and meanwhile, using AIC test to obtain p, q value. The ARIMA model of AIDS incidence was selected as ARIMA (2, 1, 4). The ARIMA model of AIDS mortality cases was selected as ARIMA (4, 1, 4). The result is shown in Figure 5. From that, we can see that the ordinary ARIMA method is challenging to deal with periodic sequences.

Therefore, the ARIMA method after the sequence decomposed is adopted. Using seasonal_decomposition function to decompose the original sequence into the form of trend sequence, seasonal sequence, residual sequence summed, the trend sequence, residual sequence were ARIMA modeling, set AIDS incidence of the trend sequence model is ARIMA (4, 1, 3), residual sequence model is ARIMA (4, 1, 3); AIDS mortality cases of the trend sequence model is ARIMA (2, 1, 4), and the residual sequence model is ARIMA (2, 1, 4). The MAE for AIDS incidence cases is 690.41, and the MAPE is 15.67%. The MAE for AIDS mortality cases is 332.18, and the MAPE is 18.32%.

## SARIMA Model

### s Parameter

Use the seasonal_decompose function to decompose the original sequence into a trend sequence, seasonal sequence, and residual sequence sum form, analyze the seasonal sequence, and set $s = 12$.

### Remaining Parameters

Use the ADF test to obtain $d$=1 and $D$=1. Apply grid-search to get $p$, $q$, $P$, and $Q$ using the AIC test. The SARIMA model of AIDS incidence was selected as SARIMA (1, 1, 1)(0, 1, 1) 12. The SARIMA model of AIDS mortality cases was selected as SARIMA (1, 1, 1)(1, 1, [1, 2]) 12. The MAE for AIDS incidence cases is 953.42, and the MAPE is 22.02%. The MAE for AIDS mortality cases is 198.1, and the MAPE is 11.38%.

## Prophet Model

According to the sequence characterizations and model feature set parameters are as follows: trend term function g(t)=logistic, followability changepoint_prior_scale=0.05, prediction type freq=ms (month output is the result of the first day of each month), yearly periodicity yearly_seasonality=1, holidays_prior_scale=10, weight holidays_prior_scale=10. The MAE for AIDS incidence cases is 707.06, and the MAPE is 16.91%. The MAE for AIDS mortality cases is 257.25, and the MAPE is 15.78%.

## BPNN Model

In the construction of the BPNN, using the series_to_supervised function to transform the original sequence into a dataset for a supervised learning task, set the parameters n_in=12, n_out=1, and finally take the 12 data of the previous year of this month as the input layer component, and the prediction result as the output layer component. Set the learning rate of the BP neural network as 0.01, the number of training times is 3000, and the number of nodes in the input, hidden, and output layers are 12, 12, and 1, respectively. The hidden layer activation function is the login function. The MAE for AIDS incidence cases is 704.89, and the MAPE is 17.65%. The MAE for AIDS mortality cases is 188.78, and the MAPE is 11.21%.

## LSTM Model

The construction of the LSTM model using a four-layer LSTM network, starting with a two-layer long short-term memory network layer with 128 neurons, an activation function that takes the Tanh function, a dropout layer that discards neurons at a dropout rate of 0.5, and a fully connected layer that uses the Rectified Linear Unit (ReLU) function as the activation function. The model optimizer selected Adaptive Moment Estimation (Adma), set the root mean square error MAE as the loss function, the number of iterations was 3000, and the time step was 12.

The original time sequence was transformed into a 165*13 series set of supervised learning tasks using the series_to_supervised function to input the model. We use the first 80% of the series as the training set and the last 20% as the prediction set. The training set is divided into training data and validation data in a ratio of 3:1. The MAE for AIDS incidence cases is 662.66, and the MAPE is 15.74%. The MAE for AIDS mortality cases is 185.46, and the MAPE is 11.19%.

## LSTM-SARIMA Combination Model

Since the LSTM model has the advantages in data following, and the SARIMA model can deal with residual sequences, we used the LSTM-SARIMA combination model in this study. Using the seasonal_decompose function to decompose the original time sequence into three sequences: trend series, periodical series, and residual series.[30] Using Min-Max

normalization on the time series and the period series, convert into the form of a supervised learning task sequence set and then input into the LSTM model and fitted by the LSTM model. Using Sigmoid normalization on the residual series and then applying the SARIMA model to search for the optimal fitting effect.

In the experiment, the SARIMA model of AIDS incidence was selected as SARIMA X(3, 1, 2) × (0, 1, [1], 12), and the AIC value (207.7) is minimal. The SARIMA model of AIDS mortality cases was selected as SARIMA X(3, 1, 2) × (0, 1, [1], 12), and the AIC value (42.4) is minimal. The LSTM-SARIMA model combined the LSTM and SARIMA models to complete the forecast. The MAE for AIDS incidence cases is 507.87, and the MAPE is 11.63%. The MAE for AIDS mortality cases is 167.62, and the MAPE is 9.94%.

## Discussion

These results show that the models ARIMA, SARIMA, and Prophet in forecasting AIDS mortality cases are higher than forecast AIDS incidence cases in accuracy. Among them, the description of sequence features is clear, with few comprehensive errors and the highest prediction adaptability in the Prophet model. The ARIMA and SARIMA of sequence decomposition have different adaptability to sequences with various features. The Predictive accuracy of deep learning models such as BPNN and LSTM is generally higher than traditional methods.[33] BPNN compared to LSTM, the forecast of LSTM is relatively more accurate. The LSTM-SARIMA model considers the advantages of the LSTM and SARIMA methods and has higher data adaptability performance, and thus, the forecast accuracy is improved.

The forecast results of the various methods are slightly different in sequences with quality characteristics. In AIDS incidence forecast, due to the large fluctuations of data and more hidden features, among the ARIMA, SARIMA, and Prophet three modeling methods, the Decomposed ARIMA model of sequence decomposition can decompose the features and perform the best. Meanwhile, the Prophet model considers multiple factors, such as holidays, and the forecast effect is better than the SARIMA method. The deep learning methods LSTM and BPNN can control the forecast accuracy within 15.5% higher than the ARIMA, SARIMA, and Prophet remaining three models. LSTM has the best prediction effect among the individual methods, and its average prediction error is only 14.31%. In the prediction of AIDS incidence number sequence, the prediction error of the LSTM-SARIMA combination method is 11.63%, which is higher than the prediction effect of LSTM alone.

The trend of the monthly AIDS mortality case sequences is less variable, and the fluctuation is smoother, so the forecast results of all models are significantly better than that of the incidence sequences. Using ARIMA, SARIMA, and Prophet three models to forecast AIDS mortality cases, the Prophet method still has good prediction performance, with a prediction deviation of 16.9%. Meanwhile, the SARIMA model performs well in the prediction of the smooth series, with a prediction deviation of 11.38%, which is lower than that of the prophet method and the ARIMA model with the sequence decomposition. LSTM and BPNN perform well in predicting the AIDS mortality cases, with the prediction errors being 11.21% and 11.19%, respectively. The prediction error of the combined method LSTM-SARIMA is 9.94%, which is an apparent improvement compared with the other models.

In conclusion, the ARIMA model does not consider the periodicity factor, and it is difficult to deal with periodic sequences, so sequences with periodicity are not suitable to be processed using the ordinary ARIMA model. In the decomposed ARIMA model, the period of its sequence is obtained by the function operation, and the ARIMA model itself only deals with the trend and residuals, thus eliminating the effect of periodicity and solving the deficiency of ordinary ARIMA in periodicity.

The inclusion of seasonal parameters in the SARIMA model gives the method the ability to handle periodic series. Compared with the ARIMA model after series decomposition, the SARIMA model has more parameters and better adjustment ability, so it has more advantages in processing residual factors of the series. The SARIMA model performs better in the smoother series of AIDS mortality cases. However, the SARIMA model has a certain lack of following the trend of the series, which leads to the high results of the method in AIDS incidence cases prediction with the trend changes significantly, and it performs less well than the ARIMA model after the series decomposition. The Prophet model, which calculates and fits trend, periodicity, holiday, and residual terms separately, is highly adaptable to traditional modeling methods and can have good prediction results in both AIDS incidence and mortality series. However, it does not consider the training set series by default when calculating the variation points, which makes it

imperfect in following the series trend and performs worse than the smooth series in the series with significant trend shifts. Also, this makes the Prophet model sometimes underperform a particular method in sequence prediction.

Compared with the traditional modeling methods, the deep learning method is more effective in tracking the trend of the sequence, which makes the prediction effect of the BPNN and LSTM models better than traditional modeling methods. BPNN can only propagate the signal in the forward direction and has a weaker ability to deal with the temporal features, which limits its application in complex environments. The LSTM model has more layers of neural networks, and the structure of gates allows the model to retain more temporal features of the past time sequences. The LSTM model has more layers, and the gate structure can make the model keep more of the past temporal features of the time series, so its prediction results are more accurate than the BPNN in more fluctuating sequences. Therefore, LSTM is generally more effective than BPNN and has the most favorable prediction performance among the individual predictions of each method.

The combined LSTM-SARIMA model preserves the high degree of tracking of sequence trends of the deep learning models and utilizes the SARIMA model to fit the residual factors of the sequences, which improves the overfitting of the residual terms caused by the over-sensitivity of the LSTM model, and thus provides better prediction results compared to the LSTM method. Using cross-validation, we can conclude that the LSTM-SARIMA model can meet the demand for predicting the number of AIDS incidence cases and mortality cases. Moreover, the LSTM-SARIMA model can significantly improve the prediction accuracy of AIDS incidence cases and mortality cases compared with the traditional modeling method.

The trend in the AIDS epidemic highlights the importance of both the design and effectiveness of AIDS prevention and control policies. By analyzing historical data on AIDS incidence and mortality, we can predict future developments in the epidemic. This predictive analysis not only provides technical support for decision-making regarding AIDS diagnosis, treatment, and medical supply distribution but also enables the evaluation of the strengths and weaknesses of current prevention and control strategies by comparing predicted outcomes with actual data. In the "Political Declaration on HIV and AIDS: Ending Inequalities and Getting on Track to End AIDS by 2030", issued on June 8, 2021, the United Nations committed to achieving the "three 95%" goals: ensuring that 95% of individuals living with HIV are diagnosed, 95% of those diagnosed receive antiretroviral therapy (ART), and 95% of those receiving treatment achieve viral suppression by 2030.[34] The prediction model proposed in this study demonstrates higher accuracy than traditional modeling methods, thereby offering enhanced support to clinicians and public health departments in their efforts to control the spread of HIV.

This study has some limitations in data collection and model design. The model's maximum performance is directly related to the data quality used. The data used in this study are the monthly AIDS data released by the Chinese CDC, which are limited to univariate only in the dimension, with lower data scale accuracy, and fail to exert the performance of deep learning fully.

Additionally, the development of AIDS is also related to various factors, including policy, social, and economic conditions. The COVID-19 pandemic in 2020 led to lockdown measures in China, which had repercussions on the spread and mortality of AIDS in the country. Unfortunately, the model did not account for this issue in its design and predictions, which may have affected its overall performance.

Therefore, to improve the sensitivity of the models, it is necessary to increase the accuracy of the models in predicting the progression of AIDS incidence cases and mortality cases in subsequent studies. It is also vital to combine it with clinical studies to find more accurate and extensive data on AIDS incidence cases and mortality cases, increase the dimensions of the data coverage, and carry out multivariate and multifactorial studies to improve the accuracy and portability of the model.

## Ethical Approval

This study was approved by the Medical Ethics Committee of Capital Medical University under the approval number 2024SY231. Permission to study data from public databases on AIDS-related official websites was obtained from the committee, and all data involved in the study were free of patient privacy.

## Acknowledgments

The authors would like to thank the staff of the Chinese Center for Disease Control and Prevention for providing AIDS data and for their efforts in the interest of public safety.

## Disclosure

The authors declare no competing interests in this work.

## References

1. Michael SG. Pneumocystis pneumonia –Los Angeles. 1981. *Am J Public Health*. 2006;96(6):980–983. doi:10.2105/ajph.96.6.980 PMID: 16714472.
2. He N, Detels R. The HIV epidemic in China: history, response, and challenge. *Cell Res*. 2005;5(11–12):825–832. PMID:16354555. doi:10.1038/sj.cr.7290354
3. UNAIDS. Global HIV & AIDS statistics – FACT SHEET [EB/OL]; 2022. Available from: https://www.unaids.org/en/resources/fact-sheet. Accessed November 29, 2023.
4. National Center for AIDS/STD Control and Prevention, China CDC. National AIDS STD epidemic, second quarter, 2023. *Chin J AIDS & STD*. 2023;29(09):953. doi:10.13419/j.cnki.aids.2023.09.01
5. Wang L, Guo W, Li D, et al. HIV epidemic among drug users in China: 1995–2011. *Addiction*. 2015;110(Suppl 1(S1)):20. doi:10.1111/add.12779
6. Wei C, Herrick A, Raymond H, et al. Social marketing interventions to increase HIV/STI testing uptake among men who have sex with men and male-to-female transgender women. *Cochrane Database Syst Rev*. 2011;(9):CD009337. doi:10.1002/14651858.CD009337
7. Acquired Immunodeficiency Syndrome and Hepatitis C Professional Group, Society of Infectious Diseases, Chinese Medical Association. Consensus on diagnosis and management of immunological non-responder in acquired immunodeficiency syndrome (version 2023). *Chin J Infect Dis*. 2024;42(1):3–13.
8. Xu B, Li J, Wang M. Epidemiological and time series analysis on the incidence and death of AIDS and HIV in China. *BMC Public Health*. 2020;20(1):1906. PMID: 33317484; PMCID: PMC7734828. doi:10.1186/s12889-020-09977-8
9. Luo Z, Jia X, Bao J, et al. a combined model of SARIMA and prophet models in forecasting AIDS Incidence in Henan Province, China. *Int J Environ Res Public Health*. 2022;19:5910. doi:10.3390/ijerph19105910
10. Zhao T, Liu H, Bulloch G, Jiang Z, Cao Z, Wu Z. The influence of the COVID-19 pandemic on identifying HIV/AIDS cases in China: an interrupted time series study. *Lancet Reg Health West Pac*. 2023;36:100755. PMID: 37360868; PMCID: PMC10072954. doi:10.1016/j.lanwpc.2023.100755
11. Wu HL, Qian JS, Xu XD, et al. BP-neural network as a model of predicting STD/AIDS prevalence. *Chin J AIDS STD*. 2007;6:525–528.
12. Li Z, Li Y. A comparative study on the prediction of the BP artificial neural network model and the ARIMA model in the incidence of AIDS. *BMC Med Inform Decis Mak*. 2020;20:143. doi:10.1186/s12911-020-01157-3
13. Wang G, Wei W, Jiang J, et al. Application of a long short-term memory neural network: a burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China. *Epidemiol Infect*. 2019;147:e194. doi:10.1017/S095026881900075X
14. Zhao Z, Zhai M, Li G, et al. Study on the prediction effect of a combined model of SARIMA and LSTM based on SSA for influenza in Shanxi Province, China. *BMC Infect Dis*. 2023;23(71). doi:10.1186/s12879-023-08025-1
15. Yawen W, Zhongzhou S, Baohu Y, Yin Y. "The application of ARIMA model and ARIMA-GRNN model in the prediction of AIDS incidence". *Chin J Disease Cont*. 2018;22(12):91–94.
16. An Q, Wu J, Meng J, Zhao Z, Bai JJ, Li X. Using the hybrid EMD-BPNN model to predict the incidence of HIV in Dalian, Liaoning Province, China, 2004–2018. *BMC Infect Dis*. 2022;22(1):102. doi:10.1186/s12879-022-07061-7 PMID: 35093010; PMCID: PMC8799978.
17. Chen Y, He J, Wang M. A hybrid of long short-term memory neural network and autoregressive integrated moving average model in forecasting HIV incidence and morality of post-neonatal population in East Asia: global burden of diseases 2000–2019. *BMC Public Health*. 2022;22:1938. doi:10.1186/s12889-022-14321-3
18. National Disease Control and Prevention Administration [EB/OL]. Available from: https://www.ndcpa.gov.cn/jbkzzx/c100016/common/list.html. Accessed October 11, 2024.
19. National health commission of the people's Republic of China [EB/OL]. Available from: http://www.nhc.gov.cn/jkj/s2907/new_list.shtml. Accessed November 11, 2024.
20. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. 5th Edition. Wiley; 2015.
21. Box GE, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis Forecasting and Control*. Oakland, California: John Wiley & Sons; 1994:238–242.
22. Ho SL, Xie M, Goh TN. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. *Comput Ind Eng*. 2002;42(2):371–375. doi:10.1016/S0360-8352(02)00036-0
23. Shibata R. Selection of the order of an autoregressive model by akaike's information criterion. *Biometrika*. 1976;63(1):117–126. doi:10.1093/biomet/63.1.117
24. Li Y, Liu X, Li X, et al. Interruption time series analysis using autoregressive integrated moving average model: evaluating the impact of COVID-19 on the epidemic trend of gonorrhea in China. *BMC Public Health*. 2023;23:2073. doi:10.1186/s12889-023-16953-5
25. Taylor SJ, Letham B. Forecasting at scale. *Peer J Preprints*. 2017;2017:1.
26. Ding M, Zhang Y. Time series prediction using series decomposition and prophet model. *Comp Syst Applicat*. 2023;32(11):294–301. doi:10.15888/j.cnki.csa.009282
27. Wang L, Zeng YR, Zhang JL, Huang W, Bao YK. The criticality of spare parts evaluating model using an artificial neural network approach. *Lect Notes Comput Sci*. 2006;3991:728–735.
28. Wang L, Zeng Y, Chen T. Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. *Expert Syst Appl*. 2015;42(2):855–863. doi:10.1016/j.eswa.2014.08.018

29. Li X, Xu X, Wang J, Li J, Qin S, Yuan J. Study on prediction model of HIV incidence based on GRU neural network optimized by MHPSO. *IEEE Access*. 2020;8:49574–49583. doi:10.1109/ACCESS.2020.2979859
30. Hochreiter S, Schmidhuber JR. Long short-term memory[J]. *Neural Comput*. 1997;9(8):1735–1780. doi:10.1162/neco.1997.9.8.1735
31. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput*. 2000;12(10):2451–2471. doi:10.1162/089976600300015015
32. Wei D, Haibin Z, Feng H, et al. Communication jamming signals recognition based on LSTM network and feature fusion [J]. *Telecommunic Engin*. 2022;62(4):450–456.
33. Zhixin Z, Xiaoxia Z, Yancen Z, Lanfang G, Liang C, Xiuyang L. Development and comparison of predictive models for sexually transmitted diseases—AIDS, gonorrhea, and syphilis in China, 2011–2021. *Front Public Health*. 2022;2022:1. doi:10.3389/fpubh.2022.966813.2296-2565.
34. Joint United Nations Programme on HIV/AIDS (UNAIDS). Political Declaration on HIV and AIDS: ending Inequalities and Getting on Track to End AIDS by 2030, UNAIDS. Available from: https://www.unaids.org/en/resources/documents/2021/2021_political-declaration-on-hiv-and-aids. Accessed October 11, 2024.