BMC Public Health

RESEARCH

Open Access

Study on the prediction performance of AIDS monthly incidence in Xinjiang based on time series and deep learning models

Dandan Tang[1,5], Yuanyuan Jin[2*], XuanJie Hu[1], Dandan Lin[3], Abiden Kapar[3], YanJie wang[3], Fang Yang[4] and Huling Li[1*]

Abstract

Objective AIDS is a highly fatal infectious disease of Class B, and Xinjiang is a high-incidence region for AIDS in China. The core of prevention and control lies in early monitoring and early warning. This study aims to identify the best model for predicting the monthly AIDS incidence in Xinjiang, providing scientific evidence for AIDS prevention and control.

Methods Monthly AIDS incidence data from January 2004 to December 2020 in Xinjiang were collected. Six different models, including the ARIMA (2,1,2) model, ARIMA (2,1,2)-EGARCH (2,2) combined model, ARIMA (2,1,2)-TGARCH (1,1) combined model, ETS (A, A, A) model, XGBoost model, and LSTM model, were used for fitting and forecasting.

Results All models were able to capture the overall trend of the monthly AIDS incidence in Xinjiang. In terms of RMSE and MAE, the ETS (A, A, A) model performed the best, achieving the smallest values. For the MAPE metric, the ARIMA (2,1,2)-TGARCH (1,1) model performed the best. Considering RMSE, MAE, and MAPE together, the ETS (A, A, A) model was the best-performing model in this study. The LSTM model also showed good predictive performance, while the XGBoost model and ARIMA (2,1,2) model performed relatively poorly.

Conclusion The ETS (A, A, A) model is the best model for predicting the monthly AIDS incidence in Xinjiang. Deep learning models (such as LSTM) have significant potential in time series forecasting. The XGBoost model and ARIMA (2,1,2) model may have limitations when handling time series data, and future improvements or combinations could enhance prediction performance.

Keywords Xinjiang region, AIDS incidence, Time series forecasting, Deep learning models

*Correspondence:
Yuanyuan Jin
jinyy33@163.com
Huling Li
lihuling@xjmu.edu.cn
[1] Medical Engineering College of Xinjiang Medical University, Urumqi 830017, China
[2] Basic Medical Science College of Xinjiang Medical University, Urumqi 830017, China
[3] College of Public Health of Xinjiang Medical University, Urumqi 830017, China
[4] Affiliated Cancer Hospital of Xinjiang Medical University, Urumqi 830011, China
[5] Institute of Medical Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi, China

Background

AIDS, also known as Acquired Immunodeficiency Syndrome, is a serious disease that threatens human life and health, caused by infection with the human immunodeficiency virus (HIV) [1]. Since the first case of AIDS was reported in the United States in June 1981, AIDS has rapidly spread to over 100 countries and regions worldwide [2]. According to reports from the World Health Organization [3], by the end of 2020, approximately 37.7 million people were living with HIV globally. In 2021 alone, 1.5 million people were newly infected with HIV, and 650,000

people died from HIV-related causes. Although significant progress has been made in controlling AIDS globally, the overall situation remains grim, with many patients still not receiving treatment. To this day, AIDS remains one of the most important and critical public health issues faced by both developed and developing countries [4]. Data on the AIDS epidemic in China shows that by the end of October 2020, there were 1.045 million reported cases of people living with HIV in China, with 112,000 new AIDS infections in that year. By the end of 2019, over 960,000 people were living with HIV in China, with nearly 160,000 new infections that year. The cumulative number of deaths reported had exceeded 310,000 [5].

AIDS is currently a major Class B infectious disease under prevention and control in Xinjiang. Its spread not only severely harms the health of the people in the region but also causes significant economic losses. Since the first case of HIV infection was reported in Xinjiang in 1995, the AIDS epidemic has spread to all counties, cities, and districts across the region [6]. An analysis of the annual incidence and mortality rates of AIDS in Xinjiang and nationwide from the 2004–2020 data from the Public Health Science Data Center of the Chinese CDC (see Fig. 1) reveals that both the annual incidence and mortality rates of AIDS in Xinjiang are higher than the national averages. This indicates that the task of AIDS prevention and control in Xinjiang faces tremendous pressure, and the overall situation remains severe.

AIDS is a legally classified Class B infectious disease in China, and since 2004, China has implemented network-based direct reporting for AIDS. Currently, commonly used models for infectious disease surveillance and early warning include the grey prediction model [7, 8], infectious disease dynamics prediction models [9–12], deep learning prediction models [13–16], and time series prediction models [17–20], with time series prediction models being one of the most widely used methods for AIDS surveillance and early warning. A review of domestic and international literature on AIDS reveals that the ARIMA time series model is the most widely applied for predicting AIDS incidence [21–25]. This study collected 17 years of monthly AIDS incidence data from Xinjiang, and applied the ARIMA model, ARIMA-EGARCH combined model, ARIMA-TGARCH combined model, ETS model, XGBoost model, and LSTM model to fit and predict the monthly AIDS incidence data for Xinjiang from 2004 to 2020. The prediction results of the six models were compared, providing valuable insights for future AIDS prediction and early warning.

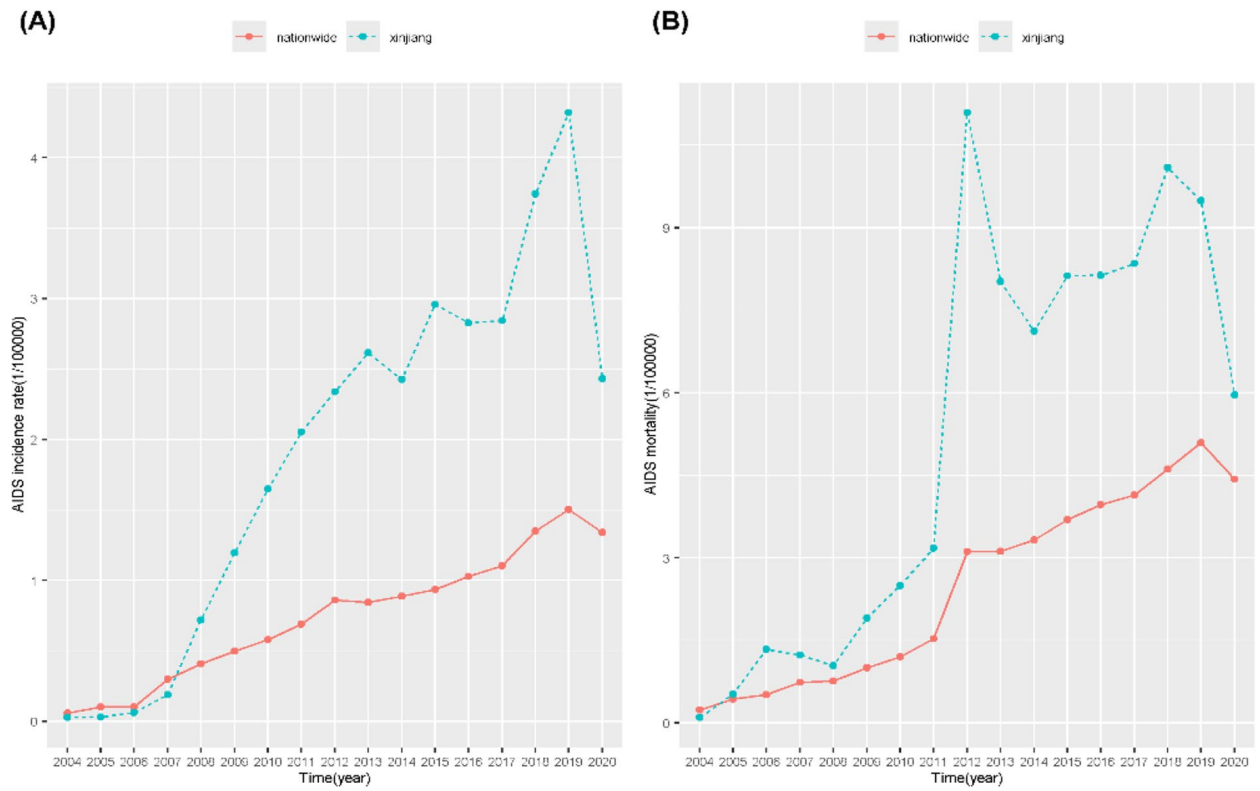

Fig. 1 **A** Trend of annual AIDS incidence in China and Xinjiang from 2004 to 2020. **B** Trend of annual AIDS mortality in China and Xinjiang from 2004 to 2020

Data Source

This study uses the monthly AIDS incidence data for Xinjiang from 2004 to 2020 obtained from the official and publicly accessible AIDS database of the Chinese Public Health Science Data Center, China Centers for Disease Control and Prevention (http://www.phsciencedata.cn/Share/en/) [26]. All data are publicly available from the web-based database, ensuring the data's accessibility and reproducibility for further research.

Data Analysis

The raw data were organized using Excel 2019. The ARIMA model was constructed using the forecast package and the Arima function in R software version 4.3.1. The ETS model was constructed using the ets function from the forecast package. The ARIMA-EGARCH combined model and ARIMA-TGARCH combined model were specified using the ugarchspec function from the rugarch package. For fitting the ARIMA model, the unit root test (augmented Dickey-Fuller, ADF) was performed using the ur.df() function from the urca package in R to check if the time series was stationary. The Ljung-Box test was applied to the residuals of the ARIMA and ETS models using the Box.test() function from the stats package in R. If the p-value was greater than 0.05, it indicated that the residuals were white noise, and the model fit was considered good.For fitting the ARIMA-EGARCH and ARIMA-TGARCH combined models, the presence of ARCH effects in the residuals of the ARIMA model was tested using the ArchTest function and the PortmanteauTest function from the FinTS package in R.The XGBoost and LSTM models were constructed using Python 3.9, with the LSTM model built using the LSTM module from the Keras library and the XGBoost model built using the XGBRegressor module from the XGBoost library.Regarding missing data, as data were missing for May, July, September, and October 2004, the na.locf() function from R was used to apply forward filling to handle the missing values.

Method

ARIMA Model

The ARIMA model, also known as the Autoregressive Integrated Moving Average model, was first proposed in the 1970s by American statistician George Box and British statistician Gwilym Jenkins as a time series analysis method. It is also referred to as the Box-Jenkins model [27]. The ARIMA model is applicable not only to stationary time series but also to non-stationary time series that have been made stationary through differencing. The ARMA model primarily includes the $MA(q)$ model, $AR(P)$ model, $ARMA(p,q)$ model, and $ARMA(p,d,q)$ model. In these models, p represents the order of the autoregressive part, d represents the number of differencing operations, and q represents the order of the moving average part. The mathematical expression for the ARIMA model is as follows:

$$\varnothing\left(B\nabla^d\right)X_t = \theta\ (B)\epsilon_t \tag{1}$$

In Formula (1): X_t represents a time series at time t, ε_t represents white noise (with zero mean and constant variance), d represents the order of differencing, and B represents the backshift operator, i.e., $BX_t = X_{t-1}, \nabla = 1 - B$. In Formula (2): $\phi\ (B)$ represents the autoregressive operator, with the autoregressive coefficient polynomial given by:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{2}$$

In formula (3), $\theta\ (B)$ represents the moving average operator, and the moving average coefficient polynomial is:

$$\theta(B) = 1 - \theta_1 B - \theta_2 B - \cdots - \phi_q B^q \tag{3}$$

TGARCH Model

Glosten, Jagannathan, and Runkle (1993) [28] proposed the Threshold Generalized Autoregressive Conditional Heteroscedasticity (TGARCH) model. The model form of TGARCH is:

$$y_t = c + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_p y_{t-p} + u_t \tag{4}$$

$$u_t \sim N\left(0, \sigma_t^2\right)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i {u_{t-i}}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 + \sum_{k=1}^{r} \alpha_k u_{t-k}^2 I_{t-k} \tag{5}$$

In formula (5): $\alpha + \beta < 1$; $\alpha \geq 0$; and $\alpha_0 > 0$; r represents the number of thresholds; I_t denotes the indicator variable; and if $u_t < 0$, then $I_t = 1$; $u_t \geq 0$, then $I_t = 0$. The conditional variance σ_t^2 is determined by both the squared residual of the previous period u_{t-1}^2 and the conditional variance σ_{t-j}^2.

EGARCH model

Nelson introduced the Exponential EGARCH model [29] to capture asymmetry. The model is expressed as:

$$y_t = \sigma_t Z_t \tag{6}$$

$$log_{\left(\sigma_t^2\right)} = \alpha_0 + \sum_{k-1}^{q} \beta_k g\left(Z_{t-k}\right) + \sum_{k-1}^{p} \alpha_k log_{\left(\sigma_{t-k}^2\right)} \tag{7}$$

$$g_{(Z_t)} = \theta Z_t + \lambda[|Z_t| - E(|Z_t|)] \tag{8}$$

$$Z_t = \frac{y_t}{\sigma_t} \tag{9}$$

In Eq. (6), σ_t^2 is the conditional variance; α_0, α, β、θ, and λ are the coefficients. Z_t can either be a standard normal variable or come from a generalized error distribution. In Eq. (8), the structure of $g_{(Z_t)}$ allows the sign and magnitude of Z_t to have different effects on volatility. Since $log_{(\sigma_t^2)}$ in Eq. (7) can be negative, the parameters are not restricted by the sign.

EST Model

The EST (Exponential Smoothing State Space Model) is a method used to predict future values by taking a weighted average of past actual observations, where more recent data is given higher weight and more distant data is given lower weight [30]. The exponential smoothing method fits the model by adding, multiplying, or applying no operation between the three main parameters: error, overall trend, and seasonality. The formula for the EST model is expressed as:

$$l_t = \alpha\left(y_t - s_{t-m}\right) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{10}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{11}$$

$$s_t = \gamma\left(y_t - l_t\right) + (1-\gamma)s_{t-m} \tag{12}$$

In formula (10), l_t represents the level at time t, y_t is the observed value, s_{t-m} is the seasonal component, and α is the smoothing parameter. In formula (11), b_t represents the trend component, and β is the smoothing parameter for the trend. In formula (12), s_t represents the seasonal component, γ is the smoothing parameter for the seasonality, and mmm is the length of the seasonal period.

XGBoost model

The XGBoost algorithm is an ensemble deep learning algorithm based on decision trees. It combines the results of all decision trees by summing them to produce the final output of the model [31], that is:

$$\widehat{y_i} = \sum_{l=\mathrm{I}}^{L} f_{l(W_i)}, f_l \mathcal{E} F \tag{13}$$

In formula (13), $\widehat{y_i}$ represents the predicted value for the i sample; f_l is the l decision tree; L is the total number of decision trees; F is the set of all decision trees; and w_i is the feature dataset of the i signal sample.

The objective function of the XGBoost algorithm is:

$$O^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t(W_i)\right) + \Omega\left(f_t\right) \tag{14}$$

$$\Omega\left(f_t\right) = \gamma T + \frac{1}{2}\lambda \sum_{\alpha=1}^{T} \omega_\alpha^2 \tag{15}$$

In Eq. (14), $\widehat{y_i}^{(t-1)}$ is the predicted value for the i sample at the t − 1 iteration; $\sum_{i-1}^{n} l\left(y_i, \widehat{y_i}\right)$ is the loss function used to measure the error between the predicted value $\widehat{y_i}$ and the true value y_i for each sample. In Eq. (15), $\Omega\left(f_t\right)$ is the regularization function that helps reduce the risk of overfitting; γ, λ are the regularization coefficients; T is the number of leaf nodes; and ω_α is the weight of the α leaf node.

LSTM model

The LSTM model was proposed by Hochreiter and Schmidhuber [32] to address the gradient explosion and gradient vanishing problems that occur in traditional recurrent neural networks. Due to its unique structure, it is widely used for modeling long time series data. The core of the LSTM network lies in the updating and passing of the cell state. The forget gate discards unimportant information from the cell state, the input gate decides what information needs to be updated in the cell state, and the output gate determines what information needs to be output.

The forget gate is given by:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{16}$$

The input is given by:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{17}$$

$$\check{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{18}$$

cell state:

$$C_t = f_t \times C_{t-1} + i_t \times \check{C}_t \tag{19}$$

The output is given by:

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{20}$$

$$h_t = O_t \times tanh(C_t) \tag{21}$$

In formulas (16–21), f_t, i_t and O_t represent the output vectors of the forget gate, input gate, and output gate at time t, respectively. W and b are the weight matrix and bias vector, with subscripts indicating the specific gate. h_t, h_{t-1} represent the hidden states at time t and t − 1, respectively. C_t and C_{t-1} are the cell state update variables at time t and t − 1, respectively. $\check{C}_t$ is the candidate cell state vector at time t.

Evaluation Metrics for Prediction Accuracy

In this study, the AIDS monthly incidence data from Xinjiang from 2004 to 2019 were used as the training set for model fitting, while the AIDS monthly incidence data from Xinjiang in 2020 (from January to December) were used as the test set to validate the model's predictions. The following evaluation metrics were chosen to assess the model's prediction performance: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) [33, 34]. These metrics were used to compare the fitting performance of different models on the training set and the predictive accuracy on the test set.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{22}$$

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{23}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2} \tag{24}$$

Results

ARIMA Model

Stationarity Test of the Series and Stationarization of Non-Stationary Series: A descriptive analysis was performed on the original monthly AIDS incidence data for Xinjiang, and a time series plot of the raw AIDS incidence data for Xinjiang from 2004 to 2020 was generated (see Fig. 2A). From Fig. 2A, it can be observed that the dataset covers the monthly AIDS incidence data from January 2004 to December 2020, with a total of 204 data points. The incidence trend can be divided into four phases. Initial Phase (2004–2007): During this period, the incidence rate data showed small fluctuations, with most months remaining between 0.01 and 0.06. Occasionally, there were higher peaks in the incidence, such as in January 2006 (0.2388) and March 2006 (0.3035), but these peaks did not form a continuous upward trend. Rising Phase (2008–2012): From 2008 onwards, the incidence data showed a gradual increasing trend, particularly between 2010 and 2012, when the incidence significantly increased. In 2012, the incidence rates were generally higher, with several months from May to December seeing rates exceeding 1, reaching a peak. Fluctuating Phase (2013–2019): During this phase, the incidence data exhibited greater volatility. From 2013 to 2015, the incidence fluctuated between 0.4 and 1. From 2016 to 2019, although there were still fluctuations, the overall incidence level decreased compared to the previous phase, with most months remaining between 0.5 and 1. Declining Phase (2020): Starting from 2020, the incidence data showed a clear downward trend. The incidence rates in 2020 were generally lower, with most months ranging between 0.1 and 0.6. Notably, in August 2020, the incidence rate sharply dropped to 0.0634, the lowest level in recent years.

Therefore, it is evident that the monthly AIDS incidence data in Xinjiang exhibits significant clustering and volatility. However, since ARIMA models are built on stationary time series, it is necessary to conduct a unit root Augmented Dickey-Fuller (ADF) test [35] on the raw AIDS incidence data for Xinjiang, with the results shown in Table 1. From Table 1, the ADF test statistic is −1.7802, and the p-value is 0.0755, which is much higher than the critical values at the 1%, 5%, and 10% significance levels. Therefore, we accept the null hypothesis (H0): the Xinjiang AIDS monthly incidence series has a

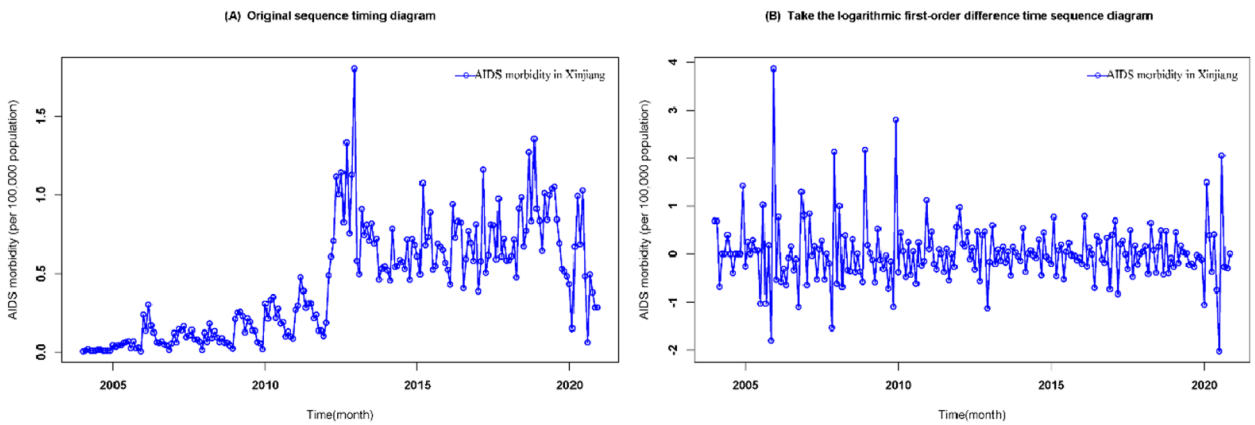

Fig. 2 Time Series Plots. **A** Time series plot of the original series. **B** Time series plot after logarithmic transformation and first differencing

Table 1 ADF test results of original sequence monthly incidence of AIDS in Xinjiang

variable	T value	*P* value
ADF Test Statistic	−1.7802	0.0755
1% Significance Level	−2.58	0.01
5% Significance Level	−1.95	0.05
10% Significance Level	−1.62	0.1

Table 2 The comparison of ADF test results

variable	T value	*P* value
Original Data	−1.7802	0.0755
Logarithmic Transformation and First-Differenced Data	−12.976	< 0.001

unit root, indicating that the series is non-stationary. For non-stationary series, stationarization can be achieved through differencing, logarithmic transformation, or a combination of both. Thus, we applied a logarithmic transformation and first-order differencing to the Xinjiang AIDS monthly incidence series, and the time series plot of the log-transformed and first-differenced series is shown in Fig. 2B. The ADF test was then performed on the transformed series, and the results were compared with the ADF test results of the original series (see Table 2). From Table 2, the ADF test statistic for the log-transformed and first-differenced AIDS monthly incidence series is −12.976, with a p-value < 0.001, which is much smaller than the critical value of −2.58 at the 1% significance level. Thus, we reject H0: the Xinjiang AIDS monthly incidence series has a unit root, indicating that the log-transformed and first-differenced Xinjiang AIDS monthly incidence series is stationary and does not contain a unit root.

ARIMA Model Identification

For the original data, logarithmic transformation and first differencing (d = 1) were applied to make the series stationary. After observing the autocorrelation plot (Fig. 3A) and the partial autocorrelation plot (Fig. 3B) of the log-transformed and differenced stationary series, and considering that a higher order may lead to overfitting, we proceeded with a stepwise approach starting from higher-order models to lower-order models. Initially, p and q were set to 0, 1, and 2, and the monthly AIDS incidence series for Xinjiang from January 2004 to December 2019 was fitted. A total of 9 ARIMA(p,d,q) candidate models were tested, with the fitting results shown in Table 3.From Table 3, it can be seen that the model with the smallest Akaike Information Criterion (AIC) [36] and Bayesian Information Criterion (BIC) [37], as well as the smallest RMSE, MAE, and MAPE, was selected as the optimal model. The final optimal model is ARIMA (2,1,2) (AIC = 322.52, BIC = 338.79, RMSE = 0.189, MAE = 0.116, MAPE = 37.595). A white noise test was conducted on the residuals of ARIMA(2,1,2), with the Ljung-Box statistic Q = 2.9434 and p-value = 0.7087 (p > 0.05), indicating that the residuals form a white noise series, and the model fits well. Therefore, the ARIMA (2,1,2) model was used to forecast the AIDS incidence rate for Xinjiang from January to December 2020.

GARCH Family Model

To better capture the short-term fluctuations in AIDS incidence and improve the prediction accuracy, we built ARIMA-EGARCH and ARIMA-TGARCH combination models based on the ARIMA(2,1,2) model. The

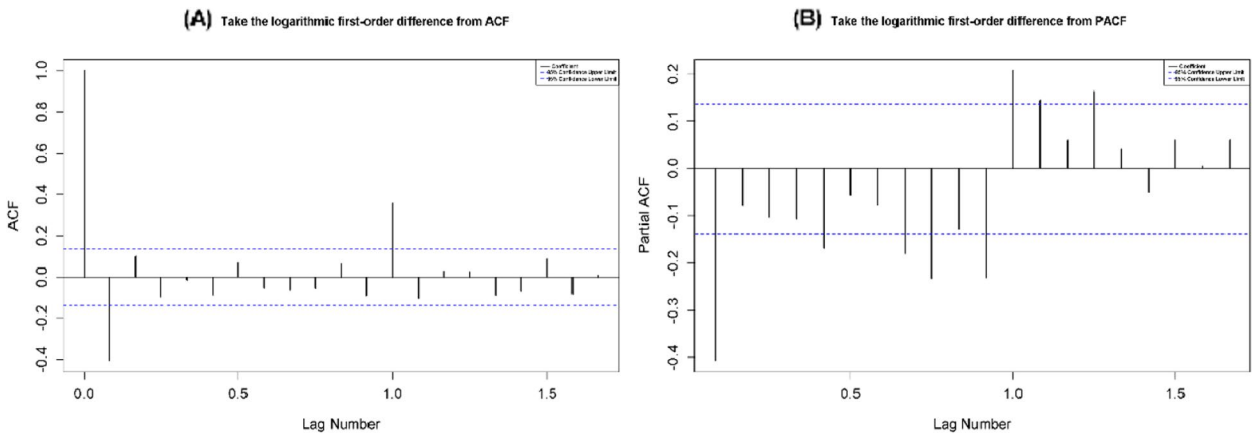

Fig. 3 Correlation Plots. **A** ACF (Autocorrelation Function) of the stationary time series. **B** PACF (Partial Autocorrelation Function) of the stationary time series

Table 3 Prediction Performance Metrics of Candidate ARIMA(p,d,q) Models on the Test Set

Model	RMSE	MAE	MAPE(%)	AIC	BIC
ARIMA (0,1,0)	0.212	0.130	40.979	364.25	367.50
ARIMA (0,1,1)	0.182	0.117	39.989	330.38	336.89
ARIMA (0,1,2)	0.182	0.117	40.310	332.32	342.08
ARIMA (1,1,0)	0.186	0.116	38.628	330.40	336.90
ARIMA (1,1,1)	0.189	0.120	39.461	328.11	337.86
ARIMA (1,1,2)	0.183	0.115	39.835	326.58	339.59
ARIMA (2,1,0)	0.185	0.116	38.785	332.32	342.08
ARIMA (2,1,1)	0.189	0.116	38.648	328.98	341.99
ARIMA (2,1,2)	0.189	0.116	37.595	322.52	338.79

Table 4 The test results of Portmanteau Q test and LM test of the *ARIMA*(2, 1, 2) model residual sequence

Order	PortmanteauQ	P	LM	P
1	23.698	1.127×10^{-6}	23.240	1.430×10^{-6}
2	24.027	6.063×10^{-6}	24.567	4.627×10^{-6}
3	24.032	2.460×10^{-5}	24.513	1.952×10^{-5}
4	24.053	7.796×10^{-5}	24.647	5.923×10^{-5}
5	24.280	0.0001918	24.568	0.0001688
6	24.963	0.0003469	24.687	0.0003902
7	25.444	0.0006328	24.578	0.0009011
8	25.702	0.001181	24.479	0.001904
9	25.784	0.002216	24.309	0.003838
10	27.304	0.002331	25.956	0.0038

optimal combination models can better address the sudden fluctuations in disease incidence, making the prediction results more reliable [38]. To test whether the residuals of the ARIMA(2,1,2) model exhibit ARCH effects, there are two common methods: the Portmanteau Q test and the Lagrange Multiplier (LM) test [39]. We performed a Portmanteau Q test and LM test with 10 lags on the residual series of the ARIMA(2,1,2) model, and the results are shown in Table 4. As shown in Table 4, the ARCH test results indicate that the p-values of the residual series from lags 1 to 10 are all less than 0.05, suggesting that the ARIMA(2,1,2) model's residual series exhibits ARCH effects. Therefore, ARIMA-EGARCH and ARIMA-TGARCH combination models can be used to fit the squared residuals of the mean model.

Considering the ease of estimating model parameters, we used the maximum likelihood method to estimate the parameters. The model order parameters p and q were selected within the range $p \leq 2$ and $q \leq 2$, resulting in 4 alternative ARIMA-EGARCH combination models and 4 alternative ARIMA-TGARCH combination models. These models were fitted to the AIDS monthly incidence data from Xinjiang, China, from January 2004 to December 2019. The fitting results of the ARIMA-EGARCH combination models are shown in Table 5, and the fitting results of the ARIMA-TGARCH combination models are shown in Table 6.

Table 5 Prediction Performance Metrics of Candidate ARIMA(2,1,2)-EGARCH(p,q) Models on the Test Set

Model	RMSE	MAE	MAPE(%)	AIC	BIC
ARIMA (2,1,2) -EGARCH (1,1)	0.189	0.116	36.578	1.329	1.499
ARIMA (2,1,2) -EGARCH (1,2)	0.199	0.122	54.438	7.687	7.873
ARIMA (2,1,2) -EGARCH (2,1)	0.191	0.117	67.454	1.362	1.566
ARIMA (2,1,2) -EGARCH (2,2)	0.194	0.117	35.686	1.240	1.461

Table 6 Prediction Performance Metrics of Candidate ARIMA(2,1,2)-TGARCH(p,q) Models on the Test Set

Model	RMSE	MAE	MAPE(%)	AIC	BIC
ARIMA (2,1,2) -TGARCH (1,1)	0.186	0.115	37.661	1.402	1.554
ARIMA (2,1,2) -TGARCH (1,2)	0.185	0.115	37.651	1.406	1.575
ARIMA (2,1,2) -TGARCH (2,1)	0.185	0.115	37.661	1.412	1.582
ARIMA (2,1,2) -TGARCH (2,2)	0.186	0.115	37.651	1.416	1.603

Based on the criteria that smaller AIC and BIC values are better, and selecting models with smaller RMSE, MAE, and MAPE values as the optimal models, Table 5 shows that, among the 4 candidate ARIMA-EGARCH combination models, the optimal model is determined to be ARIMA(2,1,2)-EGARCH(2,2) (AIC = 1.240, BIC = 1.461, RMSE = 0.194, MAE = 0.117, MAPE = 35.686). Similarly, among the 4 candidate ARIMA-TGARCH combination models, the optimal model is determined to be ARIMA(2,1,2)-TGARCH(1,1) (AIC = 1.402, BIC = 1.554, RMSE = 0.186, MAE = 0.115, MAPE = 37.661). These two models, ARIMA(2,1,2)-EGARCH(2,2) and ARIMA(2,1,2)-TGARCH(1,1), were used to predict the AIDS incidence from January to December 2020 in Xinjiang.

ETS Model

ETS (Exponential Smoothing State Space Model) is a time series forecasting method that estimates future data trends and seasonal variations by applying weighted averages to historical data. Using the training set data of AIDS monthly incidence rates in Xinjiang from January 2004 to December 2019, the following ETS models were constructed: ETS(A,N,N), ETS(A,A,N), and ETS(A,A,A). The smoothing

parameters were automatically simulated using the ets() function from the R forecast package. The optimal model was selected based on the smallest AIC and BIC values, along with relatively smaller MAPE, MAE, and RMSE values.

Through comprehensive analysis and comparison, we found that the ETS(A,A,A) model performed excellently across all evaluation metrics. Its AIC value is 744.703, BIC value is 803.338, RMSE is only 0.279, MAE is 0.143, and MAPE is as low as 37.242. These values are the smallest among the three fitted models. Therefore, we conclude that the ETS(A,A,A) model is the optimal exponential smoothing model. Based on this, we are confident in using the ETS(A,A,A) model to effectively forecast the AIDS incidence rate in Xinjiang from January to December 2020. For detailed data on the fitting results of the different ETS models, please refer to Table 7.

XGBoost Model

We used the XGBRegressor function from the XGBoost library on the Python 3.9 platform to construct the XGBoost model. The model was built using the default XGBoost parameters with some basic settings such as n_estimators = 1000 and objective = 'reg:squarederror'. This approach allowed for the construction of a well-performing XGBoost model. The best parameters for the XGBoost model and its performance metrics on the test set are shown in Table 8. Based on the data presented in Table 8, the XGBoost model effectively fitted the AIDS monthly incidence data for the Xinjiang region from January 2004 to December 2019. On the test set, the model demonstrated good performance with a Root Mean Squared Error (RMSE) of 0.318, Mean Absolute Error (MAE) of 0.195, and a Mean Absolute Percentage Error (MAPE) of 94.910. Based on these excellent performance metrics, we are confident in using this model to make effective predictions for the AIDS monthly incidence in Xinjiang from January to December 2020.

LSTM Model

Based on the Python 3.9 platform, we used the LSTM module from the Keras library to construct a Long Short-Term Memory (LSTM) model for training on the Xinjiang region's AIDS monthly incidence data from January 2004 to December 2019. During this process, several key parameters were manually set, and a high-performance LSTM model was built. The optimal parameters of the LSTM model and the prediction performance metrics on the test set are shown in Table 9. According to the data analysis in Table 9, the LSTM model demonstrated its predictive capabilities on the test set, with the following performance metrics: the Root Mean Squared Error (RMSE) was 0.443, the Mean Absolute Error (MAE) was 0.368, and the Mean Absolute Percentage Error (MAPE) was 526.311. Based on these performance metrics, we are confident that the LSTM model is capable of effectively predicting the AIDS incidence in the Xinjiang region from January to December 2020.

Table 7 Prediction Performance Metrics of Candidate ETS Models on the Test Set

Model	RMSE	MAE	MAPE(%)	AIC	BIC
ETS(A,N,N)	0.182	0.117	40.246	797.413	807.186
ETS(A,A,N)	0.183	0.119	42.575	800.470	816.758
ETS(A,A,A)	0.279	0.143	37.242	744.703	803.338

Table 8 Optimal Parameters of the XGBoost Model and Prediction Performance Metrics on the Test Set

Model	Parameter	RMSE	MAE	MAPE(%)
XGBoost model	-	0.318	0.195	94.910
booster	Gbtree			
n_estimators	1000			
max_depth	6			
learning_rate	0.3			
objective	'reg:squarederror'			
reg_alpha	0			
reg_lambda	1			

Comparison of model forecasting results

To determine the best model for predicting the monthly AIDS incidence rates in Xinjiang, we used six different time series analysis and deep learning models: ARIMA (2,1,2), ARIMA (2,1,2)-EGARCH (2,2) combined model, ARIMA (2,1,2)-TGARCH (1,1) combined model, ETS(A,A,A) model, XGBoost model, and LSTM model. These models were used to predict the AIDS monthly incidence data for Xinjiang from January to December 2020 in the test set. The trend of the predicted results is shown in Fig. 4, and the performance metrics of the predictions are presented in Table 10. As shown in Fig. 4, all six models were able to predict the AIDS monthly incidence trend in Xinjiang relatively well. According to Table 10, by calculating the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) for each model on the test set, we comprehensively evaluated the prediction performance of these models.

Comparisons of Prediction Performance Metrics: RMSE (Root Mean Square Error): The smallest RMSE value was

Table 9 Optimal Parameters of the LSTM Model and Prediction Performance Metrics on the Test Set

Parameter Category	Parameter Name	Parameter Value	RMSE	MAE	MAPE(%)
LSTM model	-	-	0.443	0.368	526.311
First Layer LSTM	units	50			
	return_sequences	True			
	input_shape	(1, 1)			
First Layer Dropout	rate	0.2			
Second Layer LSTM	units	50			
	return_sequences	False			
Second Layer Dropout	rate	0.2			
Output Layer (Dense)	units	1			
	Activation Function	Linear			
Compilation Parameters	optimizer	'adam'			
	loss	'mean_squared_error'			
Compilation Parameters	epochs	10			
	batch_size	32			

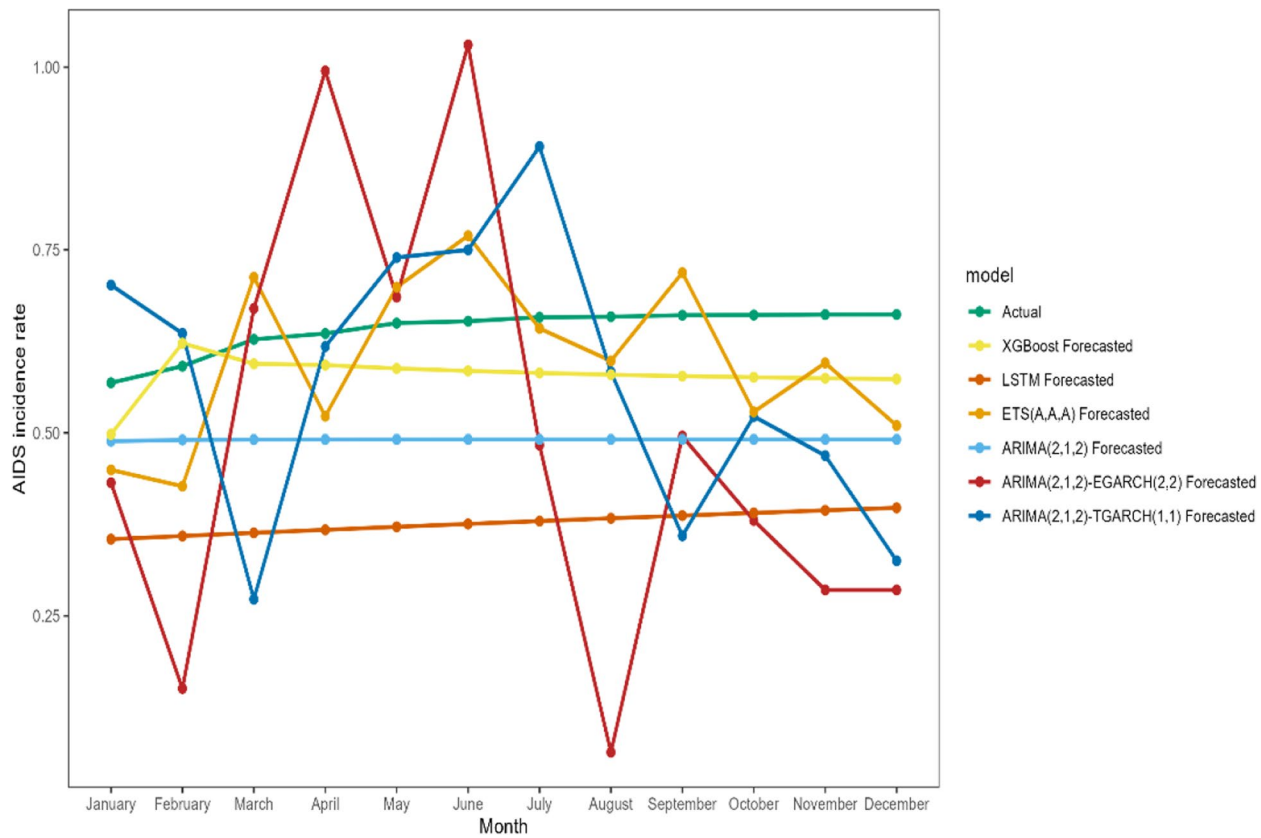

Fig. 4 Predictions of Monthly AIDS Incidence in Xinjiang for 2020 by Six Models

0.274, achieved by the ETS(A,A,A) model. The second-best was the ARIMA (2,1,2)-EGARCH (2,2) model with an RMSE of 0.301. Other models had RMSE values greater than 0.31, with ARIMA (2,1,2) at 0.324, XGBoost at 0.316, and LSTM at 0.291. MAE (Mean Absolute Error): The ETS(A,A,A) model also achieved the smallest MAE value of 0.224. The ARIMA (2,1,2)-TGARCH (1,1) model followed with an MAE of 0.246. Other models had MAE

Table 10 Comparison of the Prediction Results for Monthly AIDS Incidence in Xinjiang Using 6 Models on the Test Set

Model	RMSE	MAE	MAPE(%)
ARIMA (2,1,2)	0.324	0.280	146.168
ARIMA (2,1,2) -EGARCH (2,2)	0.301	0.252	128.509
ARIMA (2,1,2) -TGARCH (1,1)	0.318	0.246	83.562
ETS(A,A,A)	0.274	0.224	118.065
XGBoost	0.316	0.274	130.350
LSTM	0.291	0.231	103.963

values exceeding 0.25, with ARIMA (2,1,2)-EGARCH (2,2) at 0.252, XGBoost at 0.274, and LSTM at 0.231. MAPE (Mean Absolute Percentage Error): In terms of MAPE, the ARIMA (2,1,2)-TGARCH (1,1) model performed best, with a MAPE of 83.562%. Although the ETS(A,A,A) model performed excellently in RMSE and MAE, its MAPE value was 118.065%, higher than that of the ARIMA (2,1,2)-TGARCH (1,1) model. Other models had MAPE values greater than 100%, with XGBoost at 130.350%, LSTM at 103.963%, and ARIMA (2,1,2) reaching as high as 146.168%.After comprehensively considering RMSE, MAE, and MAPE, the ETS(A,A,A) model achieved the smallest values in both RMSE and MAE, indicating that the predicted values were closest to the actual values, resulting in the highest prediction accuracy. Despite performing better in MAPE, the ARIMA (2,1,2)-TGARCH (1,1) model's ranking was affected by RMSE and MAE, making it rank lower overall. The XGBoost and ARIMA (2,1,2) models performed relatively poorly. Therefore, based on the combined evaluation of RMSE, MAE, and MAPE, we can conclude that the ETS(A,A,A) model is the best-performing predictive model in this study. In summary, according to the comprehensive evaluation of RMSE, MAE, and MAPE, the ETS(A,A,A) model performed the best, followed by the ARIMA (2,1,2)-EGARCH (2,2) model. The LSTM model also demonstrated good predictive performance, while the ARIMA (2,1,2)-TGARCH (1,1) model, though the best in MAPE, ranked lower due to its RMSE and MAE. The XGBoost model and ARIMA (2,1,2) model showed relatively poor performance.

Discussion

AIDS is a highly fatal class B infectious disease caused by the HIV virus, also known as the immunodeficiency virus. Since AIDS was introduced into China in 1985, its prevalence has evolved from sporadic outbreaks to localized epidemics, and later to widespread epidemics. As of now, all 32 provinces, autonomous regions, and municipalities in mainland China have reported cases, posing a severe threat to public health. Xinjiang, in particular, is one of the high-incidence areas for AIDS in China. By September 2019, the region had 48,423 living cases, including 13,996 AIDS patients, ranking sixth in the country. Additionally, 16,004 deaths have been reported, with cases spread across the entire region. The epidemic in Xinjiang has evolved through different phases: from the sporadic phase in 1995, the localized epidemic from 1996 to 1999, and the widespread epidemic from 2000 to the present. The speed, volume, and scope of the epidemic have placed significant pressure on the region's AIDS prevention and control efforts.

One of the core strategies for AIDS prevention and control is the early identification of unusual trends and growth in AIDS cases. Monitoring, forecasting, and early warning systems are essential technologies for achieving effective AIDS prevention and control. Therefore, accurately predicting AIDS incidence is of great practical significance, providing scientific evidence for the development, improvement, and evaluation of AIDS prevention and control measures. This study aims to identify the best model for predicting the monthly AIDS incidence in Xinjiang. By comparing the performance of six different time series analysis and deep learning models–ARIMA (2,1,2), ARIMA (2,1,2)-EGARCH (2,2), ARIMA (2,1,2)-TGARCH (1,1), ETS(A,A,A), XGBoost, and LSTM–we have drawn a series of meaningful conclusions.

Based on our research, we collected monthly AIDS incidence data from Xinjiang from January 2004 to December 2020, covering a 17-year period. By analyzing the epidemiological trends of AIDS in Xinjiang, and observing the time series plot of the original AIDS incidence data (Fig. 2A), we found that the overall trend of AIDS incidence in Xinjiang exhibited significant fluctuations throughout the observation period, especially during the periods from 2012 to 2013 and from 2018 to 2020, where the incidence reached noticeable peaks. At the annual level, there were significant differences in the incidence rate across different years. For example, the incidence rate was generally higher in 2012 and 2019, while it was relatively lower in 2004 and 2008. The overall fluctuation trend of AIDS incidence in Xinjiang from 2004 to 2020 showed an upward trend followed by a downward trend, which is consistent with research results from regions like Shandong [41, 42]. This trend can be attributed to factors such as increased attention from government departments on AIDS prevention and control, efforts by public health personnel, and the relatively slow progression of the disease. During the observation period, there was a clear seasonal fluctuation in the monthly AIDS incidence rate in Xinjiang, with higher incidence rates generally observed in the spring

and winter compared to the summer and autumn. This is consistent with findings from the Xinjiang Production and Construction Corps and Hunan Province [43, 44]. Such seasonal fluctuations may be related to factors such as climate changes, patterns of human activity, and the characteristics of pathogen transmission.

We used six different time series analysis and deep learning models, including the ARIMA (2,1,2) model, ARIMA (2,1,2)-EGARCH (2,2) combined model, ARIMA (2,1,2)-TGARCH (1,1) combined model, ETS (A,A,A) model, XGBoost model, and LSTM model, to fit and predict the monthly AIDS incidence data of Xinjiang for both the training and testing sets. From the prediction trend chart (Fig. 4), it can be seen that all models were able to capture the overall trend of the monthly AIDS incidence rate in Xinjiang. This indicates that the selected models theoretically possess certain predictive capabilities, suggesting that both time series models and deep learning models are widely used in AIDS prediction research and have achieved good predictive performance [45–49]. In terms of comparing the performance metrics of the six models, RMSE, MAE, and MAPE are important indicators for measuring the accuracy of predictions and provide a comprehensive basis for model evaluation. From the RMSE metric, the ETS(A,A,A) model achieved the lowest value of 0.274, demonstrating the best performance in terms of the error between predicted and actual values. Following closely is the ARIMA(2,1,2)-EGARCH(2,2) model, with an RMSE of 0.301, also showing high prediction accuracy. This result indicates that, among traditional time series models, combined models that account for heteroscedasticity (such as EGARCH) can improve prediction performance to some extent [50]. In terms of MAE, the ETS(A,A,A) model also performed excellently, with the lowest MAE value of 0.224. Notably, the ARIMA(2,1,2)-TGARCH(1,1) model also performed well in MAE, with a value of 0.246. This suggests that considering volatility and asymmetry in time series can improve prediction accuracy to some extent. However, in the MAPE metric, we observed a different ranking. The ARIMA(2,1,2)-TGARCH(1,1) model achieved the smallest MAPE value of 83.562%, while the ETS(A,A,A) model had a relatively high MAPE value of 118.065%. This discrepancy may be due to the sensitivity of MAPE to data scale or outliers. Other researchers have also identified limitations of the MAPE metric in certain cases and proposed improvements.

Therefore, considering the three performance metrics—RMSE, MAE, and MAPE—we can draw the following conclusions: The ETS(A,A,A) model achieved the lowest values for both RMSE and MAE, indicating that its predictions are closest to the actual values, with the highest prediction accuracy. Although the ARIMA(2,1,2)-TGARCH(1,1) model performed better in terms of MAPE, we believe that RMSE and MAE are generally considered more important performance metrics, and that MAPE may be influenced by data scale or outliers. Thus, we consider the ETS(A,A,A) model to be the best performing prediction model in this study. Additionally, the LSTM model also demonstrated good predictive performance, outperforming some traditional time series models in terms of both RMSE and MAE. This finding aligns with recent research in the field of machine learning and deep learning applied to time series forecasting [51, 52], indicating that deep learning models hold significant potential in time series prediction and are worthy of further exploration and application. In contrast, the XGBoost model and the ARIMA(2,1,2) model performed relatively poorly. This may be because the XGBoost model failed to fully capture the time dependencies in the data when handling time series, and the ARIMA(2,1,2) model may not have adequately accounted for heteroscedasticity and asymmetry in the data. Therefore, future research could explore improvements or combinations of these models to enhance their prediction performance.

In summary, this study compares the predictive performance of six different time series analysis and deep learning models and concludes that the ETS(A,A,A) model is the best for predicting the monthly AIDS incidence in Xinjiang. This result not only provides an effective tool for forecasting AIDS incidence in Xinjiang but also offers new insights and directions for research in the field of time series prediction.

Conclusion

By comparing the predictive performance of six different time series analysis and deep learning models (ARIMA(2,1,2), ARIMA(2,1,2)-EGARCH(2,2), ARIMA(2,1,2)-TGARCH(1,1), ETS(A,A,A), XGBoost, and LSTM), we conclude that all models are able to capture the overall trend of AIDS monthly incidence in Xinjiang, indicating that the selected models theoretically possess certain predictive capabilities. After a comprehensive analysis of model evaluation metrics, the ETS(A,A,A) model achieved the lowest values, showing optimal performance in terms of the error between predicted and actual values, thus demonstrating the highest prediction accuracy and being identified as the best predictive model. The LSTM model also exhibited good predictive performance, with better results in RMSE and MAE than some traditional time series models, highlighting the potential of deep learning models in time series forecasting. Therefore, this study identifies the best predictive model for forecasting AIDS monthly incidence in Xinjiang, which can provide a theoretical basis for predicting and controlling the AIDS epidemic.

However, there are some limitations in this study. The models used only historical time series data to establish univariate time series models, focusing on the impact of time factors on the disease. The actual incidence of AIDS is very complex, influenced by various factors. Due to the lack of multidimensional research data, this study did not incorporate these covariates into the analysis or consider the interactions between Xinjiang's AIDS monthly incidence and other influencing factors. A more accurate prediction of the AIDS incidence trend in Xinjiang requires considering multiple factors. Future work will delve deeper into this issue, incorporating relevant covariate factors to build a predictive control model that better reflects the actual epidemic characteristics of AIDS in Xinjiang. Additionally, only six models were compared in this study, and there may be other superior models that were not included in the comparison, so the generalizability of the conclusion needs further validation. There may also be optimization potential in data preprocessing and feature selection to further improve model performance. Future research can explore introducing more advanced time series analysis and deep learning models for comparison to find better predictive models. In-depth research into data preprocessing and feature selection will be conducted to optimize data processing workflows, improve the quality of model input data, and further enhance model predictive performance.

Abbreviations

HIV Human immunodeficiency virus
AIDS Acquired immunodeficiency syndrome

Acknowledgements

This work was support by the Xinjiang Medical University High-level Talent Introduction Program grant funded by the Xinjiang Medical University,and by Dr. Kai Wang and the graduate students at Xinjiang Medical University,who provided invaluable assistance in mentoring and software learning.

Authors' contribution

Dandan Tang was responsible for research design, data analysis, and article writing; Yuanyuan Jin was responsible for research coordination and material support; Xunjie Hu was responsible for data compilation and statistical analysis; Dandan Lin was responsible for software processing and analysis; Abiden Kapar was responsible for data collection and compilation; Yanjie Wang was responsible for data compilation; Fang Yang was responsible for revealing research results; Huling Li was responsible for article guidance and revision. No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication.I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part. All the authors listed have approved the manuscript that is enclosed.

Funding

This paper was supported by the Xinjiang Medical University High-level Talent Introduction Program.

Data availability

The study used the database of the China Center for Disease Control and Prevention and the China Public Health Science Data Center (http://www.phsciencedata.cn/Share/En) The monthly incidence rate data of HIV in Xinjiang from 2004 to 2019 in the officially released and publicly available AIDS database, all of which can be obtained publicly from web-based databases, ensuring data availability and reproducible research.

Declarations

Ethics approval and consent to participate

No ethical issues were addressed in this paper, as we did not study any human or animal subjects, nor did we collect any personal information or sensitive data.
No conflict of interest exits in the submission of this manuscript, and manuscript is approved by all authors for publication.I would like to declare on behalf of my co-authors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part.All the authors listed have approved the manuscript that is enclosed.That its publication has been approved by the responsible authorities at the institution where the work carried out.

Consent for publication

The co-authors agreed to publication in "BMC Infectious Diseases".The copyright to the article is transferred to "BMC Infectious Diseases" effective if and when the article is accepted for publication. The author warrants that his/her contribution is original and that he/she has full power to make this grant. The author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors.

Competing interests

The authors declare no competing interests.

Received: 15 October 2024 Accepted: 17 February 2025
Published online: 25 February 2025

References

1. Li M, Yang HH, Liu Q, et al. A case of viral resistance in a patient with AIDS combined with lymphoma in the short term after anti-human immunodeficiency virus treatment. Chin J Infect Chemother. 2023;23(2):229–32.
2. Liu DC. A 30-year review of the AIDS epidemic and its spread in China. Emerg Infect Dis. 2017;01:50–2.
3. Kavanagh MM, Nygren-Krug H. Ending AIDS and stopping pandemics through closing inequalities. Am J Physiol Lung Cell Mol Physiol. 2021;321(6):L1055–6.
4. Zhang LF, Wang JJ. Analysis of psychological care and peer education during antiretroviral therapy for HIV patients. Chin Foreign Med care. 2018;37(12):133–6.
5. Chuai ZR, Zhang YH, Zhao YL, et al. Overview of the latest global and Chinese AIDS outbreaks. Information on infectious diseases. 2020;33(6):501–3.
6. Wang JL, Li F, Ni MJ. History of AIDS prevention and treatment in Xinjiang . Disease Prevention and Control. Bulletin. 2020;35(01):51–5+83.
7. Qin LJ, Wang T, Song YL. Prevalence characteristics and grey GM(1,1) model prediction of HIV/AIDS cases in Erqi District, Zhengzhou City, 2003–2020. Modern Disease Prevention and Control. 2023;34(8):591–5.
8. Chen P, Li F, Ye QH, et al. Residual modified GM(1,1) model predicts short-term HIV infection trends in Zhuzhou City. Henan J Prev Med. 2020;31(9):658–61+676.
9. D'Orso I, Forst CV. Mathematical models of AIDS-1 dynamics, transcription, and latency. Viruses [J]. 2023;15(10):2119.
10. Fernández MF, Distefano M, Mangano A, Sen L, Aulicino PC. Intra-host dynamics and co-receptor usage of HIV-1 quasi-species in vertically infected patients with phenotypic switch. Infect Genet Evol. 2020;78: 104066.
11. Le Hingrat Q, Sereti I, Landay AL, Pandrea I, Apetrei C. The Hitchhiker Guide to CD4+ T-Cell Depletion in Lentiviral Infection. A Critical Review of the Dynamics of the CD4+ T Cells in SIV and AIDS Infection. Front Immunol. 2021;12:695674.
12. Lin DD, Zeng T, Zhang M, et al. Predicting and analysing the dynamics of HIV transmission among men who have sex with men in Urumqi, Xinjiang Province. Chin J Infect Control [J]. 2019;18(05):388–95.
13. Pulliam L, Liston M, Sun B, Narvid J. Using neuronal extracellular vesicles and machine learning to predict cognitive deficits in AIDS. J Neurovirol. 2020;26(6):880–7.

14. Xu Y, Lin Y, Bell RP, Towe SL, Pearson JM, Nadeem T, Chan C, Meade CS. Machine learning prediction of neurocognitive impairment among people with AIDS using clinical and multimodal magnetic resonance imaging data. J Neurovirol. 2021;27(1):1–11.
15. Pluta A, Wolak T, Sobańska M, Gawron N, Egbert AR, Szymańska B, Horban A, Firląg-Burkacka E, Bieńkowski P, Sienkiewicz-Jarosz H, Ścińska-Bieńkowska A, Desowska A, Rusiniak M, Biswal BB, Rao S, Bornstein R, Skarżyński H, Łojek E. AIDS and age underlie specific patterns of brain abnormalities and cognitive changes in high functioning patients. Neuropsychology. 2019;33(3):358–69.
16. Wang B, Liu F, Deveaux L, Ash A, Gerber B, Allison J, Herbert C, Poitier M, MacDonell K, Li X, Stanton B. Predicting adolescent intervention non-responsiveness for precision AIDS prevention using machine learning. AIDS Behav. 2023;27(5):1392–402.
17. Werle JE, Teston EF, Rossi RM, Marcon SS, Sá JS, Frota OP, Ferreira Júnior MA, Andrade GKS. AIDS/AIDS and the social determinants of health: a time series study. Rev Bras Enferm. 2022;75(4): e20210499.
18. Chen J, Xu J, Zhou Y, Luo Y. AIDS detection and delayed diagnosis: a time series analysis in China. Int J Environ Res Public Health. 2022;19(24):16917.
19. Osei E, Amu H, Kye-Duodu G, Kwabla MP, Danso E, Binka FN, Kim SY. Impact of COVID-19 pandemic on tuberculosis and AIDS services in Ghana: an interrupted time series analysis. PLoS ONE. 2023;18(9): e0291808.
20. Feder AF, Pennings PS, Petrov DA. The clarifying role of time series data in the population genetics of AIDS. PLoS Genet. 2021;17(1): e1009050.
21. Mussina K, Kadyrov S, Kashkynbayev A, Yerdessov S, Zhakhina G, Sakko Y, Zollanvari A, Gaipov A. Prevalence of AIDS in Kazakhstan 2010–2020 and its forecasting for the next 10 years. AIDS AIDS (Auckl). 2023;15:387–97.
22. Xu B, Li J, Wang M. Epidemiological and time series analysis on the incidence and death of AIDS and AIDS in China. BMC Public Health. 2020;20(1):1906.
23. Ghazy RM, Al Awaidy S, Taha SHN. Trends of AIDS indicators in Egypt from 1990 to 2021: time-series analysis and forecast toward UNAIDS 90–90-90 targets. BMC Public Health. 2023;23(1):625.
24. 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(1):143.
25. Miller RL, Chiaramonte D, Strzyzykowski T, Sharma D, Anderson-Carpenter K, Fortenberry JD. Improving timely linkage to care among newly diagnosed AIDS-infected youth: results of SMILE. J Urban Health. 2019;96(6):845–55.
26. Public Health Science Data Centre HIV Database [EB/OL]. http://www.phsciencedata. cn/Share/ky_sjml.jspid=c2ca694e-3995-4c7f-9078-3ed0aaf14556,2020-2-24.
27. Claris S, Peter N. Arima model in predicting of Covid-19 epidemic for the Southern Africa region. Afr J Infect Dis. 2022;17(1):1–9.
28. Glosten LR, Jagannathan R, Runkle DE. On the relation between the expected value andthe volatility of the nominal excess return on stocks. J Financ. 1993;48(5):1779–801.
29. Ullah I, Muhammad Hasanat S, Aurangzeb K, Alhussein M, Rizwan M, Anwar MS. Multi-horizon short-term load forecasting using hybrid of LSTM and modified split convolution. PeerJ Comput Sci. 2023;9: e1487.
30. Utama IBKY, Pamungkas RF, Faridh MM, Jang YM. Intelligent IoT platform for multiple PV plant monitoring. Sensors (Basel). 2023;23(15):6674.
31. Zhao D, Zhang R. A new hybrid model SARIMA-ETS-SVR for seasonal influenza incidence prediction in mainland China. J Infect Dev Ctries. 2023;17(11):1581–90.
32. Huang Y, Ni Z, Lu Z, He X, Hu J, Li B, Ya H, Shi Y. Heterogeneous temporal representation for diabetic blood glucose prediction. Front Physiol. 2023;14: 1225638.
33. Wan Y, Song P, Liu J, Xu X, Lei X. A hybrid model for hand-foot-mouth disease prediction based on ARIMA-EEMD-LSTM. BMC Infect Dis. 2023;23(1):879.
34. Zhao R, Liu J, Zhao Z, Zhai M, Ren H, Wang X, Li Y, Cui Y, Qiao Y, Ren J, Chen L, Qiu L. A hybrid model for tuberculosis forecasting based on empirical mode decomposition in China. BMC Infect Dis. 2023;23(1):665.
35. Kębłowski P, Welfe A. The ADF–KPSS test of the joint confirmation hypothesis of Unit autoregressive root. Econ Lett. 2004;85(2):257–63.
36. Akaike H. A new look at the statistical identification problem. IEEE Trans Auto Control. 1974;19:716–23.
37. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6:461–4.
38. Xin J, Zhou J, Yang SX, Li X, Wang Y. Bridge structure deformation prediction based on GNSS data using Kalman-ARIMA-GARCH model. Sensors (Basel). 2018;18(1):298.
39. Katayama N. The portmanteau tests and the LM test for ARMA models with uncorrelated errors[M]. Advances in time series methods and applications. Springer New York. 2016;78:131–50.
40. Yuan Y. Prediction of national HIV/AIDS incidence based on ARIMA and LSTM [D]. Chongqing: Chongqing University; 2022.
41. Weihong C, Yanling Lv, Haiyun L, et al. Epidemiological characteristics and prediction analysis of hepatitis E virus infection in Yantai City, Shandong Province, from 2007 to 2021 [J]. Disease Surveillance. 2023;38(8):1–7.
42. Yuanyuan W, Fei T, Jinglei L. Application of time series analysis in predicting the incidence of HIV/AIDS cases in Dongcheng District, Beijing. Disease Surveillance. 2017;32(9):731–4.
43. Lan Ma, Wenwen Y, Xiaoling Ma, et al. Application of SARIMA model and Holt-Winters exponential smoothing model in predicting HIV/AIDS cases in Xinjiang Production and Construction Corps [J]. Journal of Preventive Medicine Information. 2024;40(11):1339–45.
44. Ying L, Wei T, Tianxiao Z, et al. Application study of the ARIMA multiplicative seasonal model in HIV infection in Hunan Province. Practical Preventive Medicine. 2018;25(6):760–3.
45. Wang G, Wei W, Jiang J, Ning C, Chen H, Huang J, Liang B, Zang N, Liao Y, Chen R, Lai J, Zhou O, Han J, Liang H, Ye L. 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.
46. Shi M, Lin J, Wei W, Qin Y, Meng S, Chen X, Li Y, Chen R, Yuan Z, Qin Y, Huang J, Liang B, Liao Y, Ye L, Liang H, Xie Z, Jiang J. Machine learning-based in-hospital mortality prediction of HIV/AIDS patients with Talaromyces marneffei infection in Guangxi, China. PLoS Negl Trop Dis. 2022;16(5): e0010388.
47. Zhan B, Wei W, Xie Z, Meng S, Bao X, He X, Xie X, Zhang M, Ye L, Jiang J, Yang S, Liang H. Machine learning-based prognostic prediction for hospitalized HIV/AIDS patients with cryptococcus infection in Guangxi, China. BMC Infect Dis. 2024;24(1):1121.
48. Zhu Z, Zhu X, Zhan Y, Gu L, Chen L, Li X. Development and comparison of predictive models for sexually transmitted diseases-AIDS, gonorrhea, and syphilis in China, 2011–2021. Front Public Health. 2022;10: 966813.
49. Albrijawi MT, Alhajj R. LSTM-driven drug design using SELFIES for target-focused de novo generation of HIV-1 protease inhibitor candidates for AIDS treatment. PLoS ONE. 2024;19(6): e0303597.
50. Liu X, Yue FJ, Guo TL, Li SL. High-frequency data significantly enhances the prediction ability of point and interval estimation. Sci Total Environ. 2024;912: 169289.
51. Singh V, Khan SA, Yadav SK, Akhter Y. Modeling global monkeypox infection spread data: a comparative study of time series regression and machine learning models. Curr Microbiol. 2023;81(1):15.
52. Nan Y, Gao Y. A machine learning method to monitor China's AIDS epidemics with data from Baidu trends. PLoS ONE. 2018;13(7): e0199697.

Publisher's Note