scientific reports

Check for updates

OPEN

Prediction analysis of human brucellosis cases in Ili Kazakh Autonomous Prefecture Xinjiang China based on time series

Lian Lu[1,5], Tongxia Yang[2,5], Zhisheng Chen[3], Qidi Ge[3], Jing Yang[4]✉ & Gan Sen[1]✉

Human brucellosis remains a significant public health issue in the Ili Kazak Autonomous Prefecture, Xinjiang, China. To assist local Centers for Disease Control and Prevention (CDC) in promptly formulate effective prevention and control measures, this study leveraged time-series data on brucellosis cases from February 2010 to September 2023 in Ili Kazak Autonomous Prefecture. Three distinct predictive modeling techniques—Seasonal Autoregressive Integrated Moving Average (SARIMA), eXtreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks—were employed for long-term forecasting. Further, the optimal model will be used to explore the impact of COVID-19 on the transmission of Human brucellosis in the region. We constructed a SARIMA(4,1,1)(3,1,2)12 model, an XGBoost model with a time lag of 22, and an LSTM model featuring 3 LSTM layers and 100 neurons in the fully connected layer to predict monthly reported cases from January 2021 to September 2023. The results indicated that the occurrence of brucellosis exhibits pronounced seasonal patterns, with higher incidence during summer and autumn, peaking in June annually. Performance evaluations revealed low Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE) for all three models. Specifically, the coefficient of determination (R^2) was 0.6177 for the SARIMA model, 0.8033 for the XGBoost model, and 0.6523 for the LSTM model. The study found that the XGBoost model outperformed the other two in long-term forecasting of brucellosis, demonstrating higher predictive accuracy. This discovery can aid public health departments in advancing the deployment of prevention and control resources, particularly during peak seasons of brucellosis. It was also found that the impact of the COVID-19 pandemic on the transmission of human brucellosis in the region was minimal. This research not only provides a reliable predictive tool but also offers a scientific basis for formulating early prevention and control strategies, potentially reducing the spread of this disease.

Keywords Brucellosis, Brucella, SARIMA, XGBoost, LSTM

Human brucellosis is a typical zoonosis caused by Brucella bacteria, which includes *Brucella melitensis* found in goats and sheep, *Brucella abortus* in cattle, *Brucella canis* in dogs, and *Brucella suis* in pigs[1]. This pathogen is a gram-negative, tiny, spherical bacterium that grows exclusively inside cells and has a tenacious ability to survive under extreme environmental conditions such as temperature, humidity, and pH[2]. Brucellosis is primarily transmitted through direct contact with secretions (such as aborted fetuses, placentas, or other bodily fluids) from infected animals, including cattle, sheep, goats, and pigs. Consumption of unpasteurized dairy products or contaminated food and water can also lead to infection. Inhalation of aerosols containing Brucella, though rare, is also possible. In animals, the disease typically manifests as abortions, stillbirths, premature births, and postpartum metritis, with infected female animals discharging bacterial secretions that serve as a source of infection. In humans, early symptoms of infection include fever, general fatigue, joint pain, and night sweats. If left untreated, the disease can lead to various complications involving the cardiovascular and nervous systems, and in severe cases, can be life-threatening[3–5]. Although advancements in medical technology have reduced the

[1]Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830017, China. [2]The Second People's Hospital of Yining, Yining 835000, China. [3]Ili Kazak Autonomous Prefecture Center for Disease Control and Prevention, Yining 835000, China. [4]Ili Friendship Hospital, Yining 835000, China. [5]Lian Lu and Tongxia Yang have contributed equally to this work. ✉email: ; sengan99@163.com

global incidence of brucellosis, the incidence remains high in some developing countries with limited medical resources[6].

In China, brucellosis remains an urgent public health challenge[7,8]. According to data from the China Public Health Data Center (http://www.phsciencedata.cn/Share/), the incidence rate of human brucellosis in the Xinjiang Uyghur Autonomous Region (referred to as Xinjiang) is 16.32 per 100,000, highlighting the severe situation faced by the region. Particularly in the Ili Kazakh Autonomous Prefecture (referred to as Ili) in Xinjiang, the incidence rate of human brucellosis has shown an upward trend, ranking first among all regions in Xinjiang. This trend poses a serious threat to the stable development of local animal husbandry and the social economy. Therefore, there is an urgent need to develop accurate epidemiological prediction models to effectively respond to this public health challenge. Considering China's vast geographical area and significant differences in climate, environment, and economic development levels among provinces (autonomous regions), it is particularly crucial to formulate targeted disease prevention and control strategies.

Accurately predicting the number of zoonotic brucellosis cases and their trends is crucial for public health departments to promptly identify high-risk areas and populations, as well as to implement effective preventive measures to halt disease transmission. Additionally, it aids in the rational allocation of medical resources, timely deployment of healthcare personnel and supplies, and the conduct of targeted health education and promotional activities. The prediction methodologies primarily encompass statistical prediction, machine learning prediction, and deep learning prediction.

Statistical prediction methods involve the application of mathematical and statistical principles to forecast future events or trends. These methods include time series analysis and regression analysis. Time series analysis focuses on capturing time dependence in data, utilizing models such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Generalized Linear Autoregressive Moving Average (GLARMA). These models are well-suited for analyzing and predicting data that exhibits trends, seasonality, and other temporal patterns. In infectious disease prediction research, the SARIMA model is widely applied[9–14]. As a time series analysis model[15–17], SARIMA can handle non-stationary time series, and capture trends, seasonality, and other linear information within the data, thereby enabling modeling and prediction.

In contrast, machine learning methods, including neural networks and deep learning, are more effective in extracting nonlinear information. As a subset of machine learning, deep learning utilizes multi-layer neural networks and excels at capturing complex nonlinear relationships within data. XGBoost, a machine learning method, has applications across various domains[18–21] but is relatively limited in its application to infectious disease prediction[22–24]. Specifically, in the context of brucellosis prediction, related research remains scarce.

Within the realm of neural networks, Recurrent Neural Networks (RNN) are commonly used for processing and predicting long-sequence data. The connections between nodes in the hidden layers after the input layer form loops, enabling the network to effectively process sequential data, ultimately producing results in the output layer. Long Short-Term Memory (LSTM) networks[25] have emerged as a prominent time series prediction method in the field of deep learning in recent years. Compared to RNN, LSTM addresses the issues of gradient explosion and gradient vanishing[26], thus performing exceptionally well in time series prediction and frequently applied in industrial sectors[27–30]. In the field of infectious diseases, LSTM models are primarily used for predicting diseases such as hand-foot-and-mouth disease[31], COVID-19[32–34], and dengue fever[35], but their application in brucellosis prediction remains relatively rare.

This study aims to utilize traditional SARIMA models, XGBoost models, and LSTM models to accurately predict the monthly incidence rate of brucellosis in Ili Kazakh Autonomous Prefecture, Xinjiang, China, over an extended period. The objective of the research is to evaluate the performance of these models and identify the most suitable model for predicting brucellosis in this region, thereby providing crucial early warning information to support local prevention and control efforts. Simultaneously, the optimal model is used to explore the impact of the spread of human brucellosis in the region during the COVID-19 pandemic.

Specifically, this study focuses on long-term forecasting, which entails predicting trends beyond one year into the future. In the context of time series forecasting, short-term forecasting typically refers to predicting trends within the next few weeks to several months; medium-term forecasting covers trends from several months up to one year; whereas long-term forecasting, as is the focus here, predicts trends beyond one year, aiming to provide a scientific basis for developing long-term prevention and control strategies.

Methods

Data sources

This study has received formal approval from the Ili Kazakh Autonomous Prefecture Center for Disease Control and Prevention. The data used in the research come from the county-level administrative units of the Ili Kazakh Autonomous Prefecture, covering the monthly reported cases of human brucellosis from January 2010 to September 2023. The data from January 2010 to January 2021 are used as the training set for model training, while the data from February 2021 to September 2023 are used as the prediction set to evaluate the predictive performance of the models. The diagnostic criteria for human brucellosis follow the health industry standard of the People's Republic of China, "Brucellosis Diagnosis" (WS269-2019).

Based on China's prevention and control policies for COVID-19, the time series is divided into the COVID-19 pandemic period and the non-COVID-19 period. This division starts in January 2020, when China first issued control measures for COVID-19, and ends in January 2023, when the Chinese government announced the adjustment of COVID-19 management from "Class B, Category A" to "Class B, Category B". During this specific period, it is coded as "1", while the remaining periods are coded as "0", representing either the absence of the pandemic or the "Class B, Category B" period. Subsequently, this period classification is introduced as a new variable into the optimal model to once again predict the number of human brucellosis cases.

SARIMA model

The incidence characteristics of human brucellosis are characterized by long-term trends, periodicity, and seasonality. ARIMA (Autoregressive Integrated Moving Average model) is a well-established model widely used in time series analysis, highly regarded for its long history of application and excellent performance. Its general form is ARIMA (p, d, q), where p represents the autoregressive order (AR), d represents the differencing order, and q represents the moving average order (MA). However, when time series data exhibits pronounced seasonal patterns, the traditional ARIMA model becomes insufficient. To address this issue, the SARIMA (Seasonal Autoregressive Integrated Moving Average model) was introduced, with its general form being SARIMA (p, d, q) (P, D, Q)s. This model, (p, d, q) handles non-seasonal information. In contrast (P, D, Q) handles seasonal information, where P represents the seasonal autoregressive order (SAR), D represents the seasonal differencing order, Q represents the seasonal moving average order (SMA), and S represents the seasonal period length. By incorporating seasonal components, the SARIMA model can better capture and predict time series data with distinct seasonal characteristics.

In order to deeply understand the characteristics and patterns of the data and to more accurately establish and apply the SARIMA model for future numerical predictions, we first used the STL(Seasonal and Trend decomposition using Loess) method[36,37] to decompose the time series data of human brucellosis. The STL method includes two types of decomposition: additive model decomposition and multiplicative model decomposition. In the multiplicative model decomposition, a logarithmic or Box-Cox transformation of the original series is required. In this study, we chose the additive model decomposition, whose mathematical relationship is expressed as follows:

$$Y_t = T_t + S_t + R_t$$

where $\mathrm{Y_t}$ represents the actual observed value of human brucellosis at time t, this value can be decomposed into long-term trend information $\mathrm{T_t}$, seasonal trend information $\mathrm{S_t}$, and random fluctuation information $\mathrm{R_t}$. To ensure the stability of the time series, based on the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), an Augmented Dickey-Fuller (ADF) test is employed to determine whether the entire time series is stationary. If the time series is stationary, no pre-processing is required; otherwise, preprocessing is necessary. In addition, with the assistance of grid search, we can select the optimal combination of model parameters (p, d, q, P, D, Q) according to the Akaike Information Criterion (AIC). Finally, the Ljung-Box (LB) test is used to perform a white noise test on the residuals of the model to ensure that the predictive results of the model are highly accurate.

XGBoost model

XGBoost is an additive model based on the idea of boosting ensemble, first proposed by Tianqi Chen et al[38] in 2016. After continuous research and improvement by many scientists, it has become a widely used model in multiple fields. The core concept of the XGBoost model is to achieve the best fit of the data by constructing a combination of multiple predictive functions and assigning weights to each function. The model uses the second-order Taylor expansion as the loss function, effectively addressing the complexity of calculating derivatives for some first-order loss functions. Additionally, XGBoost introduces a regularization term in the loss function, which helps prevent overfitting and improves the generalization ability of the model. The objective function of the XGBoost model is as follows:

$$\mathrm{Obj_t} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{L}((\mathrm{y_i}, \mathrm{y_i^{(t-1)}}) + \mathrm{f_t(x_i)}) + \Omega(\mathrm{f_t})$$

$$\Omega\left(\mathrm{f}_t\right) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{\mathrm{j}=1}^{\mathrm{T}} \omega_{\mathrm{j}}^2$$

where n represents the total number of samples,x_i denotes the i-th sample, $f_t(x_i)$ represents the prediction for the i-th sample,y_i signifies the observed value,$y_i^{(t-1)}$ indicates the prediction for the i-th sample x_i on the (t-1)-th tree,$\Omega(f_t)$ represents the complexity of the t-th tree,L denotes the loss function used to measure the discrepancy between predictions and actual values, and is also used for the output of new trees.γT is the L1 regularization term aimed at reducing model complexity, while $\frac{1}{2}\lambda \sum_{\mathrm{j}=1}^{\mathrm{T}} \omega_{\mathrm{j}}^2$ is the L2 regularization term used to smooth parameters and prevent overfitting.T indicates the number of leaves in the decision tree, and ω represents the weight parameters for each leaf node. To ensure the reliability of the model, we use grid search to determine specific hyperparameters, including learning rate, number of estimators, maximum depth, minimum child node weight, gamma, subsample rate, column sample rate, L1 regularization term, and L2 regularization term. Additionally, we employ GridSearchCV for grid searching parameters, with cross-validation utilizing the TimeSeriesSplit from the machine learning library sklearn. TimeSeriesSplit is designed specifically for cross-validation of time series data. It ensures that future data is not used to predict past events during model training, thereby preventing data leakage—the parameter n_splits = 5, indicating a fivefold cross-validation. By manually adjusting the time lag and determining the final value, we adjust as many model parameters as possible to enhance predictive performance and reduce loss bias, enabling the XGBoost model to effectively capture complex patterns in the data.

LSTM model

LSTM networks are a special type of recurrent neural network that can efficiently handle and predict time series data. The structure of an LSTM model consists of four parts: an input layer, LSTM layer, a fully connected layer, and an output layer. Before model training, data is usually normalized for pre-processing to reduce the differences between data points and accelerate the model's convergence. In this study, the maximum-minimum normalization method was employed. This method scales the data to the range of [-1,1], enhancing the model's robustness and training efficiency. The formula for this normalization is as follows:

$$x^* = \frac{x - min(x)}{max(x) - min(x)}$$

After determining the strategy for data normalization, it is necessary to adjust the hyperparameters of the model, including the learning rate, the number of LSTM layers, the number of neurons in the fully connected layer, and the optimizer, among others. These require careful tuning because these parameters have a crucial impact on the performance of the model. In addition, we have also implemented Dropout to prevent overfitting of the LSTM model. However, the complexity of the model is not always better. An overly complex model may lead to a decrease in generalization ability, meaning the model performs poorly on unseen data. Conversely, an overly simple model may struggle to converge, leading to unstable training and failing to achieve ideal prediction results. Therefore, parameter adjustment requires repeated trials and optimization to ensure that the model maintains good performance while also possessing sufficient generalization ability.

During the training process of a model, selecting the appropriate activation function and loss function is crucial. This study employed the gradient descent algorithm for training the model, using historical data from previous years to train the model to predict future trends in the data. The mathematical expression of the LSTM model is as follows, with its structural diagram shown in Fig. 1 $output_i = F(W_i \cdot A_i + b_i)$

$$C_t = C_{t-1} \cdot f_t + i_t \cdot C_t\prime$$

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

$$C_t\prime = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

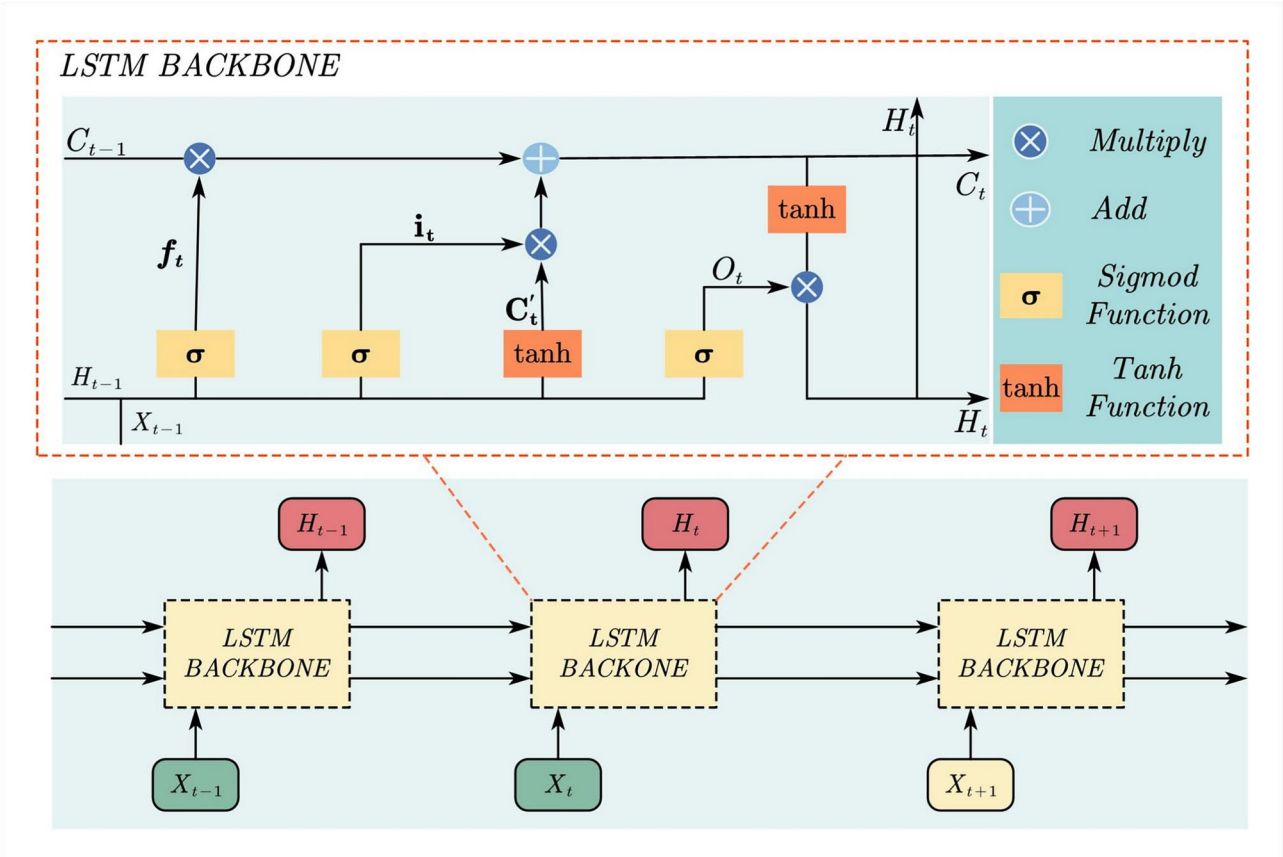

Fig. 1. LSTM model structure diagram.

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

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

$$h_t = tanh\left([C_{t-1} \cdot f_t + i_t \cdot C_t\prime] \cdot W_t + b_t\right) \cdot O_t$$

In each layer function of a neural network, there are two crucial attributes: the weight vector and the bias b. These parameters together determine the output of the neural network. The activation function represents each component of the input vector i and denotes the time.

Evaluation and comparison of models

In evaluating the model's prediction results, we used four metrics: Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Square Error (RMSE), and Coefficient of Determination (R2). Among these, the smaller values of MAE, SMAPE, and RMSE indicate that the model's prediction results are more accurate. Conversely, the closer the value of R2 is to 1, the better the model's fitting effect. The definitions of these metrics are as follows:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|x_{t,ture} - x_{t,predicted}|$$

$$SMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|x_{t,true} - x_{t,predicted}|}{(|x_{t,true} + x_{t,predicted}|)/2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x_{t,true} - x_{t,predicted}\right)^2}$$

$$\mathrm{R2} = 1 - \frac{\sum_{t=1}^{n}\left(\mathrm{x_{t,ture}} - \mathrm{x_{t,predicted}}\right)^2}{\sum_{t=1}^{n}\left(\mathrm{x_{t,ture}} - \mathrm{x_{mean}}\right)^2}$$

Data analysis

The Python language as well as relevant third-party libraries were used for writ the model. In this study, the statistical significance level was set at $\alpha = 0.05$.

Results

Analysis of the onset trend and distribution characteristics

Figure 2 illustrates the temporal trend of reported human brucellosis cases in the Ili, Xinjiang, China from January 2010 to September 2023. During this period, a total of 11,693 cases were reported. The incidence of human brucellosis from January 2010 to January 2012 was relatively low and showed a stable trend. Starting from 2012, the number of cases showed an upward trend, reaching its peak in 2015. After that, the number of cases began to decline, but the upward trend was resurgent after 2020. Further analysis from Fig. 3 shows that human brucellosis in the Ili is mainly concentrated in Yining County and Huocheng County, exhibiting a clear seasonal and periodic pattern, with a peak in cases from April to September each year, with the highest number of cases in May.

SARIMA model prediction analysis.

The time series of brucellosis cases in the Ili Kazakh Autonomous Prefecture, Xinjiang, China, is plotted in Fig. 2. It is evident from the graph that the brucellosis cases exhibit clear cyclical and seasonal characteristics. To analyze these features more thoroughly, we applied an additive model decomposition based on the Seasonal Local Trend (SLT) method to the time series data, separating the long-term trend, seasonal trend, and residual information, as shown in Fig. 4. Since the SARIMA model requires the input data to be a stationary time series, we performed a unit root test, such as the ADF (Augmented Dickey-Fuller) test, on the entire time series to confirm its stationarity. The results of the ADF test, as shown in Fig. 5(a), indicate that the original series is non-stationary, necessitating a differencing operation to achieve stationarity. After a first-order differencing operation, the P-value of the ADF test significantly decreased, approaching 0, suggesting that the series has become stationary, as shown in Fig. 5(b). Based on the results of the first-order differencing operation, we can determine that the values of d and D in the SARIMA (p, d, q) (P, D, Q)S model are both 1. Next, by observing the autocorrelation and partial autocorrelation plots, we can determine the values of p, q, P, and Q. However, this image-based method may introduce subjective judgments. Therefore, we employed a grid search method combined with the Akaike Information Criterion (AIC) to obtain the optimal values of parameters p, q, P, and Q. After multiple trials and adjustments, we found that the SARIMA (4, 1, 1) (3, 1, 2,)$_{12}$ model had the best fitting effect, with an AIC value of 1436.80. To validate the effectiveness of the model, we conducted the Ljung-Box (LB) test, which revealed that the P-value of Q was greater than 0.05, indicating that the model's residual lags are white noise. Additionally, we plotted the residual graph, as shown in Fig. 6, which further confirmed that the model's residual lags are white noise. Thus, the parameters in the SARIMA (4, 1, 1) (3, 1, 2,)$_{12}$ model were confirmed to be optimal. Finally, we used this model for prediction, with the results shown in Fig. 7(a).

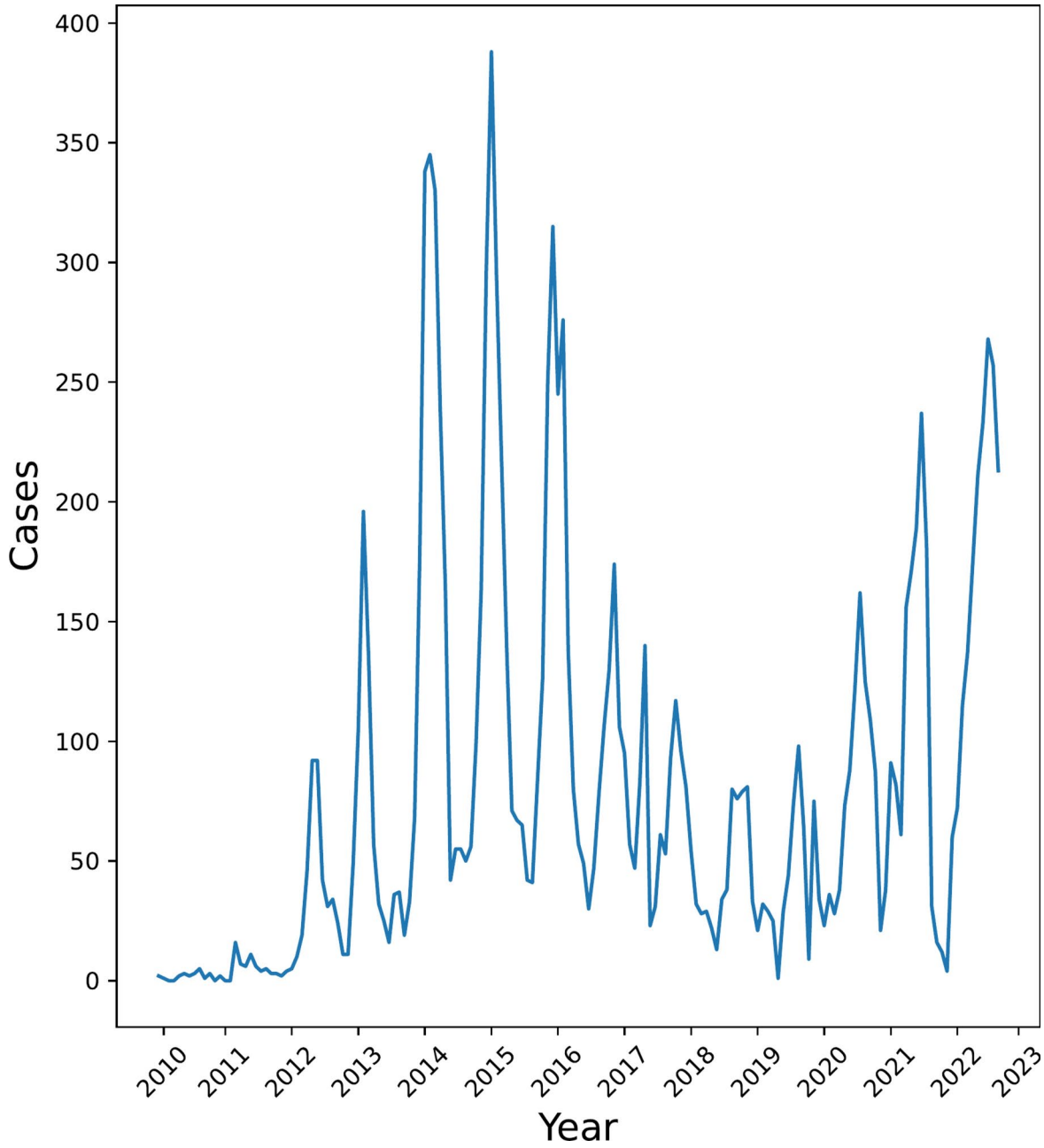

Fig. 2. Time series of human brucellosis cases in Ili, Xinjiang, China, 2010–2023.

XGBoost model prediction analysis

To accurately determine the optimal model parameters, this study employed the GridSearchCV method, combining grid search techniques to pinpoint specific parameter values. Subsequently, the optimal model parameters were further determined by manually testing the adjacent values of specific parameters. To ensure the high reliability and robustness of the prediction model, we utilize GridSearchCV for parameter grid search and adopt TimeSeriesSplit for five-fold time-series cross-validation. Specifically, TimeSeriesSplit divides the training set into five consecutive periods, with the data from each period serving as the test set for one validation round, while the data from all previous periods constitute the training set. This approach ensures that the data in the training set for each validation round temporally precedes the data in the test set, thus avoiding the use of future information to predict past situations. Ultimately, we collect and average the validation results from the five-time periods to obtain an overall assessment of the model's performance.

After a series of optimizations and validations, it was found that the performance was optimal when the lag was set to 22. To provide more detailed information, we have included the results varying with the time lag as supplementary material; see Fig S1 in the supplementary material. The optimal hyperparameters were set as follows: learning rate (learing_rate) at 0.01, the number of estimators (n_estimators) at 50, maximum depth (Max_depth) at 4, minimum child weight (min_child_weight) at 5, gamma at 78, subsample rate at 0.45, column

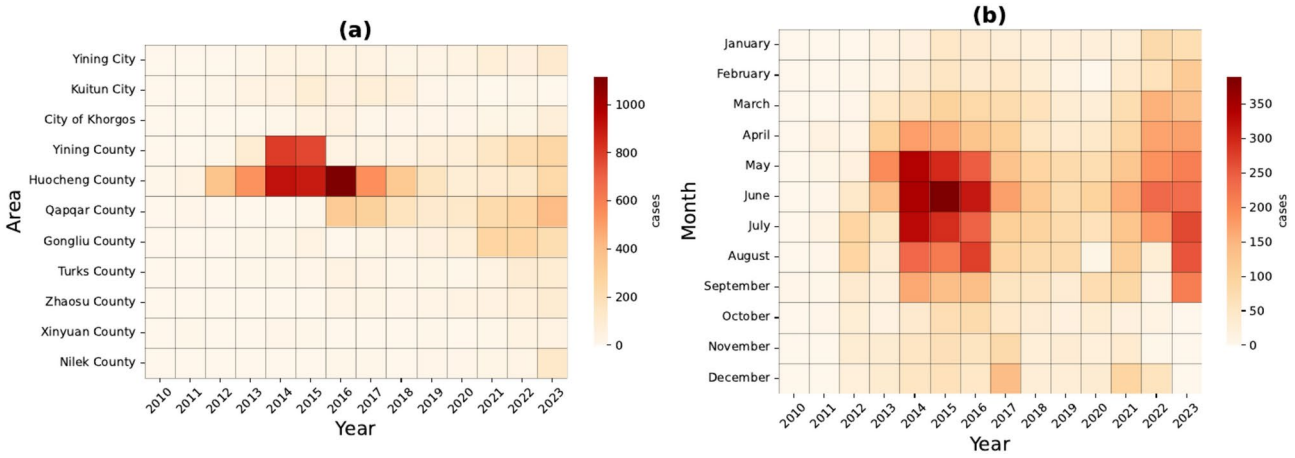

Fig. 3. Spatio-temporal aggregation of the number of reported cases of brucellosis in Ili Kazakh Autonomous Prefecture, Xinjiang, 2010–2023. (**a**) indicates the distribution of brucellosis in different counties and cities of Ili during 2010–2023. (**b**) indicates the distribution of brucellosis by month during 2010–2023.

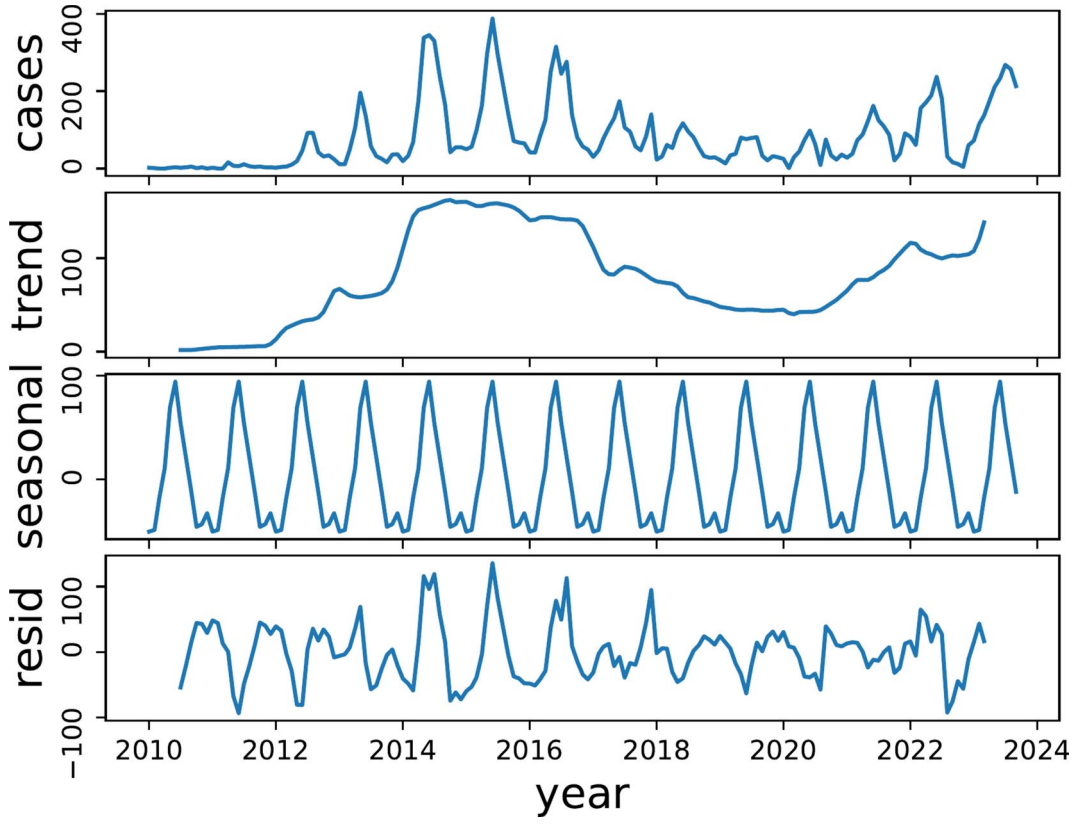

Fig. 4. Seasonal and trend decomposition of human brucellosis in Yili Kazakh Autonomous Prefecture from 2010 to 2023 based on STL.

sampling rate (colsample_bytree) at 1, L1 regularization term (reg_alpha) at 6.3, and L2 regularization term (reg_lambda) at 1. The prediction results are shown in Fig. 7(b).

LSTM model prediction analysis

In the initial phase of the experiment, we attempted to train the data using the sliding window method (by sliding a fixed-length window across the time series to generate a series of input–output pairs), but the prediction results did not meet expectations and were gap compared to the traditional method of feeding data into the model point by point (where the time series data is input into the model one time point at a time, considering only

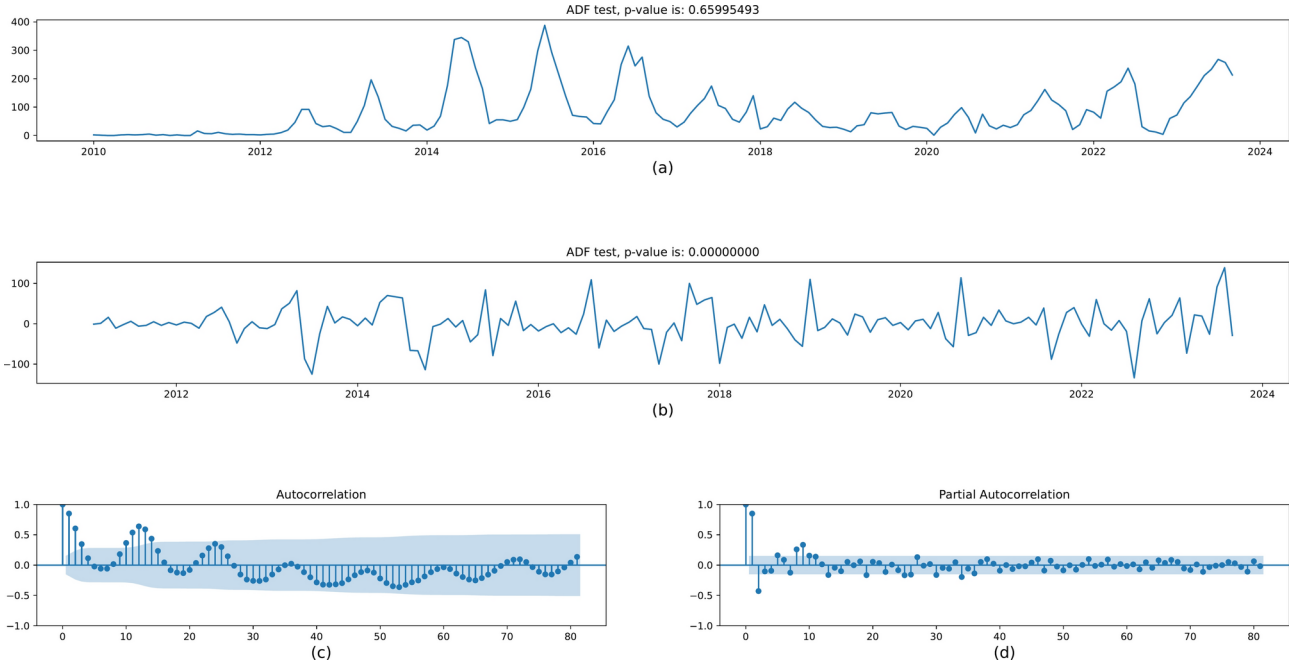

Fig. 5. ADF test before and after difference of the time series as well as autocorrelation and partial autocorrelation after difference. Where (**a**) is the sequence before first-order difference, (**b**) is the sequence after first-order difference, (**c**) and (**d**) are the autocorrelation and partial autocorrelation after difference, respectively.

the data of the current time point). Therefore, this study ultimately adopted the method of feeding the model with data point by point for training. Through repeated trials and optimization, we successfully constructed the best model, with hyperparameters set as follows: a learning rate of 0.01, 3 LSTM layers, 100 neurons in the fully connected layer, using the tanh function as the activation function, MSELoss as the loss function, Adam as the optimizer, and 3000 training epochs. The final model's loss value, loss, was 0.00002468, achieving good performance. The prediction results are shown in Fig. 7(c).

Comparison of models

Table 1 presents a performance comparison of the SARIMA, XGBoot, and LSTM models on a 32-month prediction set, including four evaluation metrics: MAE, SMAPE, RMSE, and R2.

In constructing the SARIMA (4, 1, 1)(3, 1, 2,)$_{12}$ model, since the series was not stationary, we performed first-order differencing and seasonal differencing, resulting in the exclusion of the first 13 months of data, leaving 120 months for subsequent model training. Additionally, the time lag for the XGBoost model was set to 22 months, so 111 months of data were used for model training. Therefore, all the metrics in Table 1 are based on the 32-month test set data. In terms of MAE, SMAPE, and RMSE, the performance of the SARIMA and LSTM models was significantly higher than that of the XGBoost model, indicating that the SARIMA and LSTM models had poorer prediction accuracy. Conversely, in the Coefficient of Determination (R2) metric, the performance of the SARIMA and LSTM models was significantly lower than that of the XGBoost model, reflecting their poorer fitting capabilities. Considering these metrics together, the XGBoost model is more suitable for brucellosis prediction in the Ili Kazakh Autonomous Prefecture, Xinjiang, China. The high prediction accuracy and good fitting capabilities of the XGBoost model are of great significance for the early prevention and control of brucellosis in this region. We have decided to use the XGBoost model to explore the impact of the COVID-19 pandemic on the spread of brucellosis in the region. The results are presented in Table 2.

Discussion

According to the analysis in this study, in the Ili Kazakh Autonomous Prefecture, Xinjiang, China, from 2010 to 2012, the number of human brucellosis cases remained relatively stable. However, from 2012 to 2015, there was a significant increase in the number of cases. This increase may be attributed to the improvement of the local economic situation and the expansion of the livestock industry, as well as the lack of awareness of brucellosis among residents[39]. Additionally, the relative lack of experience in brucellosis prevention and control at that time may also have been one of the reasons for the increase in case numbers. In the mid-phase of the time series, with the strengthening of publicity and awareness education on brucellosis by relevant disease control centers due to the previous prevalence of brucellosis, residents' awareness of brucellosis increased, leading to an increase in reported cases. This may have been one of the reasons why the number of cases began to decline after 2015. From 2015 to 2019, with the standardization of livestock management and the enhancement of public health awareness, the prevalence of brucellosis showed a clear downward trend. However, since 2019, the number

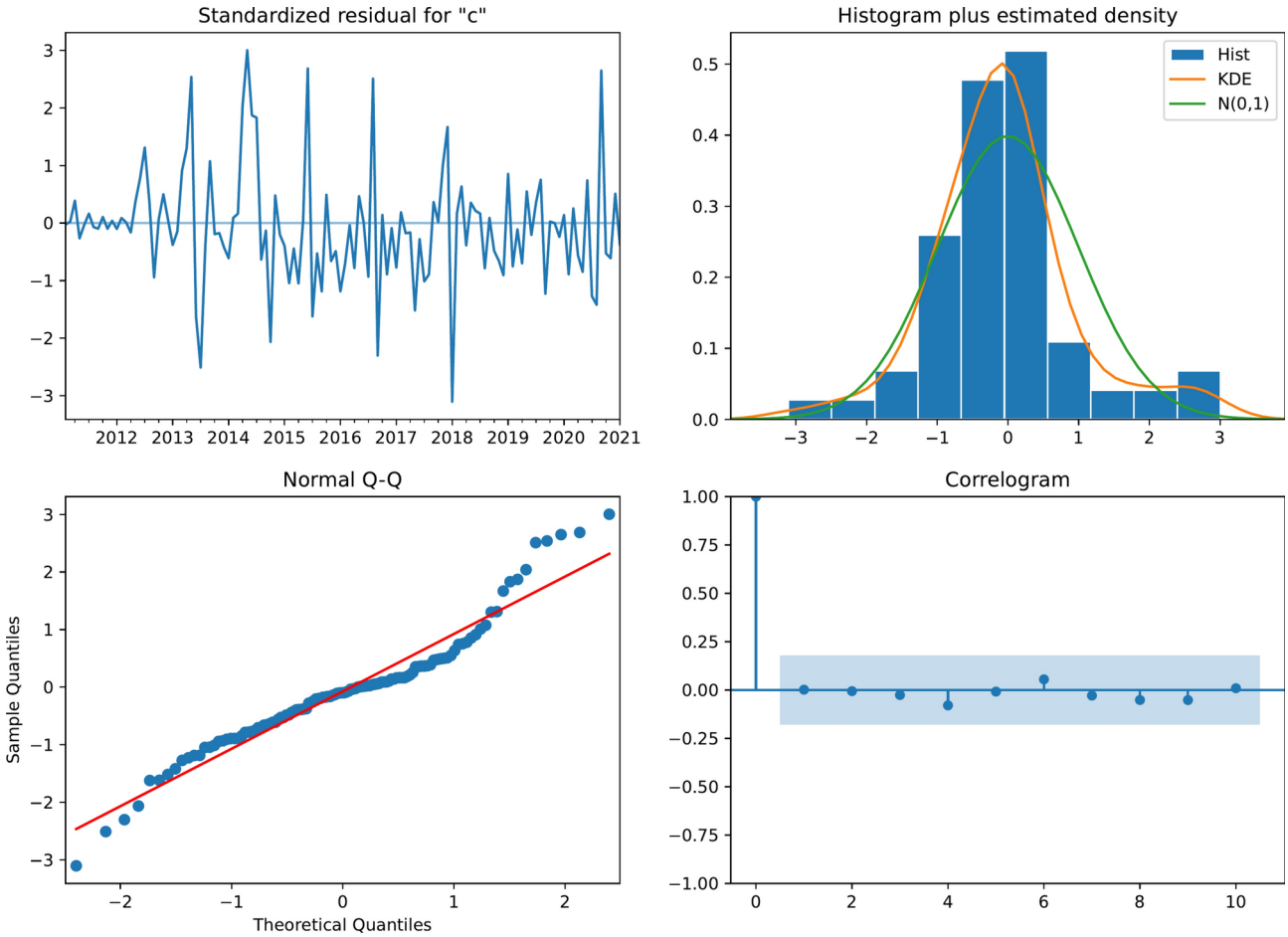

Fig. 6. Residual graph, residual histogram, standard Q, correlation.

Model	MAE	SMAPE(%)	RMSE	R2
SARIMA	36.8125	23.4204	46.9734	0.6177
XGBoost	26.4062	19.2896	33.6912	0.8033
LSTM	32.9062	20.8329	44.7957	0.6523

Table 1. Comparison of model performance: MAE, SMAPE,RMSE, R2.

Model	MAE	SMAPE(%)	RMSE	R2
XGBoost	26.4062	19.2896	33.6912	0.8033
XGBoost + CVIOD-19	26.4062	19.2896	33.6912	0.8033

Table 2. A comparison of the optimal XGBoost model indicators between the best model XGBoost and the optimal XGBoost model with the new variable "whether during the COVID-19 period" was added.

of brucellosis cases has unexpectedly shown a year-on-year upward trend, a phenomenon that deserves high attention.

Further research has found that the geographical distribution of brucellosis has also undergone significant changes. In the past, brucellosis cases were mainly concentrated in Yining County and Huocheng County, but in recent years, they have begun to spread to other counties and cities, and this trend is still intensifying, as shown in Fig. 3(a). This change may be closely related to the ongoing economic development and the increased demand for beef and lamb among residents[40,41]. With the expansion of breeding scales, the frequency of human-animal contact has increased, especially during activities such as lambing, slaughtering cattle and sheep, and shearing sheep, which increases the opportunity for exposure to Brucella[42]. At the same time, the wider circulation of animal products has also increased the risk of residents contracting brucellosis[43]. These findings provide

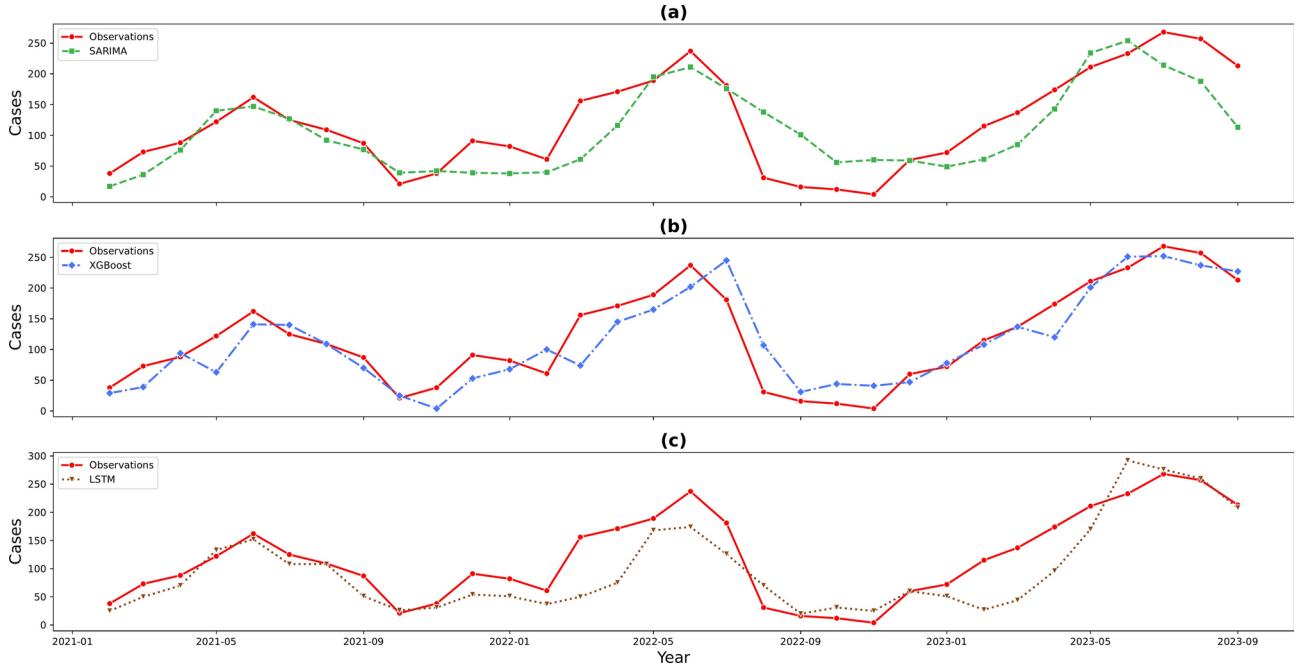

Fig. 7. Observations versus model predictions, **(a)**, **(b)**, **(c)** are SARIMA, XGBoost, LSTM model predictions versus observations respectively.

important perspectives on the transmission and prevention of brucellosis and emphasize the necessity of further strengthening disease control and public health intervention measures. According to experience in prevention and control, enhancing public awareness of brucellosis and standardized management in the livestock industry can effectively suppress the occurrence of brucellosis[44]. Additionally, for the changing geographical distribution, active detection of animal products should be strengthened to contain infected products at the early stage of circulation.

Fig. 3(b) shows that human brucellosis has higher incidence rates in summer and autumn, while it is relatively low in spring and winter, which fully demonstrates its distinct seasonal characteristics. However, this seasonal variation is different from the results of other studies[45,46] , and it may be closely related to the local climate conditions of the Ili Kazakh Autonomous Prefecture. For instance, the region experiences a short summer and a long winter, with late autumn and early spring often being affected by cold air, leading to a sharp drop in temperature, which may affect the activity of Brucella[47]. Moreover, the link between human brucellosis and the livestock industry cannot be overlooked. Sheep are the primary source of infection for this disease[40], and the Ili Kazakh Autonomous Prefecture is an important base for raising fine-wool sheep in China. Compared to the cold of early spring and winter, as temperatures rise, the biological activity and reproductive capacity of Brucella gradually recover[45]. Additionally, pastoralists begin grazing activities at this time. Since the intake and weight gain of cattle and sheep are slower, and their immune systems are weaker, the risk of brucellosis infection increases[47]. These analyses provide deeper insights, which can help in formulating more effective strategies for disease prevention and control. For example, interventions can be conducted in late spring and during the peak of infection in summer[44], or vaccination of the main infected livestock can be implemented to reduce the occurrence and spread of brucellosis.

To effectively prevent, control, and eliminate brucellosis, it is also necessary to further enhance early warning technologies. The prediction of infectious disease trends is a crucial aspect of disease control and eradication. The aim of this study was to provide a useful reference for the prevention and control of human brucellosis in the Ili Kazakh Autonomous Prefecture, Xinjiang. The SARIMA model, an extension of the ARIMA model, incorporates seasonal factors, trend factors, cyclic factors, and random errors into the time series variables. The SARIMA model is not restricted by data types, is adaptable, and can handle many infectious disease prediction tasks[47,48]. In this study, we selected SARIMA (4, 1, 1) $(3, 1, 2)_{12}$ as a benchmark for comparison with other models. However, the results showed that the predictions of this model were not ideal, with an R2 value of only 0.6177. The possible reason is that SARIMA is a statistical model and may struggle to handle non-linear information in time series. In most cases, the ability to handle non-linear information in the data is crucial, whether in the field of infectious diseases or other areas. Despite this, as seen in Fig. 7(a), the SARIMA model's predictions for the first 9 months were relatively accurate, indicating that SARIMA is more suitable for short-term predictions[49].

Compared to the SARIMA model, neural networks are well-suited for handling non-linear information in time series. In this study, the LSTM model exhibited lower MAE and RMSE than the SARIMA model, with an R2 value of 0.6523, which is superior to the R2 value of the SARIMA model, aligning with the research conclusions of Ashutosh Kumar Dubey and others[50]. For input consisting solely of time variables, the LSTM model did

not demonstrate its expected predictive capabilities. As can be seen from Fig. 7(c), the LSTM model achieved satisfactory results in the first 9 months of predictions[49].

In contrast to other models, XGBoost has demonstrated remarkable performance, especially when it comes to processing non-linear data. In the process of building the model, XGBoost can automatically choose and combine features, which improves prediction accuracy. In addition, regularization terms added during optimization reinforce the model's generalization ability. As a deep learning model, LSTM relies heavily on a large amount of data and high-quality data preprocessing to fully utilize its predictive potentia[51]. In this study, the XGBoost model demonstrated lower MAE, SMAPE, and RMSE values compared to the other two models. Moreover, its coefficient of determination R2 reached a high of 0.8033, far exceeding the R2 values of the LSTM and SARIMA models, indicating that the XGBoost model has high predictive accuracy. This conclusion is consistent with other research[52,53].

After a comprehensive comparison, it can be concluded that the XGBoost model is the best choice among the three models. In the application of brucellosis prediction, research on the XGBoost model is relatively scarce. For example, Alim et al[54] designed a national-scale model but did not consider the differences in climate, dietary habits, and other factors between different regions. This study focuses on multiple counties and cities within the same region, which have smaller differences in various aspects and thus can be considered under similar conditions. Additionally, the prediction set in this study spans 32 months and still maintains good prediction results, which is not present in other studies. Additionally, the XGBoost model offers a rich set of tunable hyperparameters, providing a basis for subsequent model tuning. Compared to the complex modeling processes of SARIMA and LSTM models, their construction is more straightforward and can be quickly applied to real-world scenarios. Therefore, the potential application of this model to the prediction of brucellosis deserves further exploration.

The indicators in Table 2 show that the impact of the COVID-19 period on the transmission of brucellosis in Yili Kazak Autonomous Prefecture, Xinjiang, China, is minimal. The prediction results of the XGBoost model, with the COVID-19 period as a new variable, are consistent with those without incorporating this variable, indicating that the COVID-19 period has a negligible effect on the transmission of brucellosis in the region. Additionally, we conducted a correlation analysis to further verify this point, with detailed information provided in Supplementary Material Fig S2. This conclusion is inconsistent with Ma C et al.[55], which focused solely on the COVID-19 period, whereas our study encompasses both COVID-19 and non-COVID-19 periods. Furthermore, thanks to China's government's prevention and control policies, the number of infections in the region during the pandemic was low, resulting in minimal impact from the COVID-19 pandemic.

Although this study offers valuable insights, it also acknowledges several limitations. Firstly, the research focuses solely on the Ili Kazak Autonomous Prefecture in Xinjiang, China. Given China's vast expanse and significant regional variations in geography, climate, and socio-economic conditions, these factors may considerably influence the prevalence of human brucellosis. Secondly, the study did not consider external factors such as cattle and sheep slaughter volumes, inventory levels, and dairy production. Incorporating these factors into the analysis could potentially enhance the accuracy of brucellosis case predictions in the region.

The purpose of this study is to provide support to healthcare workers in the Ili Kazak Autonomous Prefecture of Xinjiang for their efforts in brucellosis prevention and control, as well as to assist in disease control in similar regions. It is noteworthy that the predictions presented in this study are based on aggregated data at the county and city levels, and therefore, specific predictions at the individual county or city level are not available, which constitutes a limitation of the study.

Conclusions

In this study, we constructed SARIMA, XGBoost, and LSTM models and used them to predict brucellosis in the Ili Kazakh Autonomous Prefecture of Xinjiang. The results showed that the SARIMA model and LSTM model performed poorly in predicting the monthly number of human brucellosis cases in Xinjiang, while the XGBoost model showed excellent performance. The impact of the COVID-19 period on the transmission of human brucellosis in the region is minimal. Using XGBoost to predict brucellosis in the Ili Kazakh Autonomous Prefecture is of great significance for early warning, prevention, and control.

Data availability

As the data belong to the Center for Disease Control and Prevention of Ili Kazakh Autonomous Prefecture in Xinjiang, China, the original dataset is not publicly available during the current study period, but can be obtained from the corresponding authors upon reasonable request.

Received: 24 March 2024; Accepted: 19 November 2024
Published online: 07 January 2025

References

1. Chen, Y. et al. Changes of predominant species/biovars and sequence types of Brucella isolates, inner Mongolia. *China. BMC Infect Dis* **13**, 514 (2013).
2. Abu Sulayman, Sulaiman Mohammed et al (2020) Brucellosis: Current Status Of The Disease And Future Perspectives. Postępy Mikrobiologii - Advancements of Microbiology 59 : 337 - 344. https://doi.org/10.21307/PM-2020.59.4.25
3. Kaan, J. A. et al. Clinical manifestations and hazards of brucellosis in the Netherlands. *Ned. Tijdschr. Geneeskd.* **156**(12), A4460 (2012).
4. Golshani, M., & Buozari, S. (2017). A review of Brucellosis in Iran: Epidemiology, Risk Factors, Diagnosis, Control, and Prevention. Iranian biomedical journal, 21(6), 349–359. https://doi.org/10.18869/acadpub.ibj.21.6.349
5. Zheng, R. et al. A Systematic Review and Meta-Analysis of Epidemiology and Clinical Manifestations of Human Brucellosis in China. *Biomed. Res. Int.* **2018**, 5712920. https://doi.org/10.1155/2018/5712920 (2018).

6. Franc, K. A. et al. Brucellosis remains a neglected disease in the developing world: a call for interdisciplinary action. *BMC Public Health* **18**, 125. https://doi.org/10.1186/s12889-017-5016-y (2018).
7. Jiang, H., O'Callaghan, D. & Ding, J. B. Brucellosis in China: history, progress and challenge. *Infect Dis Poverty* **9**, 55. https://doi.org/10.1186/s40249-020-00673-8 (2020).
8. Liang, P. F., Zhao, Y., Zhao, J. H., Pan, D. F. & Guo, Z. Q. Human distribution and spatial-temporal clustering analysis of human brucellosis in China from 2012 to 2016. *Infectious diseases of poverty* **9**(1), 142. https://doi.org/10.1186/s40249-020-00754-8 (2020).
9. Wang, H., Tian, C. W., Wang, W. M. & Luo, X. M. Time-series analysis of tuberculosis from 2005 to 2017 in China. *Epidemiology and infection* **146**(8), 935–939. https://doi.org/10.1017/S0950268818001115 (2018).
10. Zhao, D. et al. Research on hand, foot and mouth disease incidence forecasting using hybrid model in mainland China. *BMC Public Health* **23**, 619. https://doi.org/10.1186/s12889-023-15543-9 (2023).
11. Wang, Y., Xu, C., Wang, Z. & Yuan, J. Seasonality and trend prediction of scarlet fever incidence in mainland China from 2004 to 2018 using a hybrid SARIMA-NARX model. *PeerJ* **7**, e6165. https://doi.org/10.7717/peerj.6165 (2019).
12. Zhao, D., Zhang, H., Cao, Q., Wang, Z. & Zhang, R. The research of SARIMA model for prediction of hepatitis B in mainland China. *Medicine* **101**(23), e29317. https://doi.org/10.1097/MD.0000000000029317 (2022).
13. Sharafi, M., Ghaem, H., Tabatabaee, H. R. & Faramarzi, H. Forecasting the number of zoonotic cutaneous leishmaniasis cases in south of Fars province, Iran using seasonal ARIMA time series method. *Asian Pacific journal of tropical medicine* **10**(1), 79–86. https://doi.org/10.1016/j.apjtm.2016.12.007 (2017).
14. Zeng, Q. et al. Time series analysis of temporal trends in the pertussis incidence in Mainland China from 2005 to 2016. *Scientific reports* **6**, 32367. https://doi.org/10.1038/srep32367 (2016).
15. Chikobvu Delson, Sigauke Caston. (2012). Regression-SARIMA modelling of daily peak electricity demand in South Africa. Journal of Energy in Southern Africa. 23. 23–30. https://doi.org/10.17159/2413-3051/2012/v23i3a3169
16. Farsi, M. et al. Parallel genetic algorithms for optimizing the SARIMA model for better forecasting of the NCDC weather data. *AEJ - Alexandria Engineering Journal* https://doi.org/10.1016/j.aej.2020.10.052 (2020).
17. Pan, S., Zheng, Z., Guo, Z. & Luo, H. An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. *Journal of Petroleum Science and Engineering.* **208**, 109520. https://doi.org/10.1016/j.petrol.2021.109520 (2021).
18. Keshavamurthy, Ravikiran, et al. "Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches." One Health 15 (2022): 100439.
19. Xiong, X., Guo, X., Zeng, P., Zou, R. & Wang, X. A Short-Term Wind Power Forecast Method via XGBoost Hyper-Parameters Optimization. *Frontiers in Energy Research* https://doi.org/10.3389/fenrg.2022.905155 (2022).
20. Ye, M. et al. Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. *The Science of the total environment* **858**(Pt 1), 159798. https://doi.org/10.1016/j.scitotenv.2022.159798 (2023).
21. Dong, W., Huang, Y., Lehane, B. M. & Ma, G. XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. *Automation in Construction* **114**, 103155. https://doi.org/10.1016/j.autcon.2020.103155 (2020).
22. Lv, C. X., An, S. Y., Qiao, B. J. & Wu, W. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC infectious diseases* **21**(1), 839. https://doi.org/10.1186/s12879-021-06503-y (2021).
23. Hou, N. et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *Journal of translational medicine* **18**(1), 462. https://doi.org/10.1186/s12967-020-02620-5 (2020).
24. Meng, D., Xu, J. & Zhao, J. Analysis and prediction of hand, foot and mouth disease incidence in China using Random Forest and XGBoost. *PloS one* **16**(12), e0261629. https://doi.org/10.1371/journal.pone.0261629 (2021).
25. Yong, Yu., Si, X., Changhua, Hu. & Zhang, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput* **31**(7), 1235–1270. https://doi.org/10.1162/neco_a_01199 (2019).
26. Alex Sherstinsky (2020) Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena, Volume 404,132306,ISSN 0167–2789. https://doi.org/10.1016/j.physd.2019.132306.
27. Zhao, J., Mao, X. & Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control.* **47**, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035 (2019).
28. Kim, T.-Y. & Cho, S.-B. Predicting Residential Energy Consumption using CNN-LSTM Neural Networks. *Energy* https://doi.org/10.1016/j.energy.2019.05.230 (2019).
29. Wu, Di & Jiang, Zhongkai & Xie, Xiaofeng & Wei, Xuetao & Yu, Weiren & Li, Renfa. (2019). LSTM Learning With Bayesian and Gaussian Processing for Anomaly Detection in Industrial IoT. IEEE Transactions on Industrial Informatics. PP. 1–1. https://doi.org/10.1109/TII.2019.2952917
30. Song, X. et al. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *Journal of Petroleum Science and Engineering* **186**, 106682. https://doi.org/10.1016/j.petrol.2019.106682 (2020).
31. Zhang, R. et al. Comparison of ARIMA and LSTM in Forecasting the Incidence of HFMD Combined and Uncombined with Exogenous Meteorological Variables in Ningbo, China. *International journal of environmental research and public health* **18**(11), 6174. https://doi.org/10.3390/ijerph18116174 (2021).
32. Ayyoubzadeh, S. M. et al. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR public health and surveillance* **6**(2), e18828. https://doi.org/10.2196/18828 (2020).
33. Tomar, A. & Gupta, N. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *The Science of the Total Environment* **728**, 138762–138762. https://doi.org/10.1016/j.scitotenv.2020.138762 (2020).
34. Kafieh, R. et al. COVID-19 in Iran: Forecasting Pandemic Using Deep Learning. *Computational and mathematical methods in medicine* **2021**, 6927985. https://doi.org/10.1155/2021/6927985 (2021).
35. Mussumeci, E. & Codeço Coelho, F. Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression. *Spatial and spatio-temporal epidemiology* **35**, 100372. https://doi.org/10.1016/j.sste.2020.100372 (2020).
36. Cleveland RB, Cleveland WS, McRae JE, Terpenning IJJoOS. STL: a seasonal-trend decomposition. J Off Stat. 1990;6(1):3–73
37. Sanchez-Vazquez, M. J., Nielen, M., Gunn, G. J. & Lewis, F. I. Using seasonal-trend decomposition based on loess (STL) to explore temporal patterns of pneumonic lesions in finishing pigs slaughtered in England, 2005–2011. *Preventive veterinary medicine* **104**(1–2), 65–73. https://doi.org/10.1016/j.prevetmed.2011.11.003 (2012).
38. Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785–794. https://doi.org/10.1145/2939672.2939785.
39. Zhai, M. et al. Research on the predictive effect of a combined model of ARIMA and neural networks on human brucellosis in Shanxi Province, China: a time series predictive analysis. *BMC Infect Dis.* **21**(1), 280 (2021).
40. Yang, H. et al. Epidemiological Characteristics and Spatiotemporal Trend Analysis of Human Brucellosis in China, 1950–2018. *International Journal of Environmental Research and Public Health.* **17**(7), 2382. https://doi.org/10.3390/ijerph17072382 (2020).
41. Peng, C. et al. Spatial-temporal distribution of human brucellosis in mainland China from 2004 to 2017 and an analysis of social and environmental factors. *Environ Health Prev Med* **25**, 1. https://doi.org/10.1186/s12199-019-0839-z (2020).
42. Pereira CR, Cotrim de Almeida JVF, Cardoso de Oliveira IR, Faria de Oliveira L, Pereira LJ, et al. (2020) Occupational exposure to Brucella spp.: A systematic review and meta-analysis. PLOS Neglected Tropical Diseases 14(5): e0008164. https://doi.org/10.1371/journal.pntd.0008164
43. Jiang, S. L. et al. The investigation report about the prevention and control of brucellosis in China. *Chinese Journal ofControl ofEndemic Diseases* **21**(6), 3–6 (2006).

44. Karyn,A,Havas,et al.The human–animal interface of domestic livestock management and production and its relationship to brucellosis in the country of Georgia 2010: A rapid assessment analysis[J].
45. Chen, H. et al. Driving role of climatic and socioenvironmental factors on human brucellosis in China: machine-learning-based predictive analyses. *Infect Dis Poverty* **12**, 36. https://doi.org/10.1186/s40249-023-01087-y (2023).
46. Wang, Y. et al. Temporal trends analysis of human brucellosis incidence in mainland China from 2004 to 2018. *Sci Rep* **8**, 15901. https://doi.org/10.1038/s41598-018-33165-9 (2018).
47. Cong, J., Ren, M., Xie, S. & Wang, P. Predicting Seasonal Influenza Based on SARIMA Model, in Mainland China from 2005 to 2018. *Int. J. Environ. Res. Public Health* **16**, 4760. https://doi.org/10.3390/ijerph16234760 (2019).
48. Qiu, H. et al. Forecasting the incidence of mumps in Chongqing based on a SARIMA model. *BMC Public Health* **21**, 373. https://doi.org/10.1186/s12889-021-10383-x (2021).
49. Liu, X., Lin, Z. & Feng, Z. Short-term offshore wind speed forecast by seasonal ARIMA - A comparison against GRU and LSTM. *Energy* https://doi.org/10.1016/j.energy.2021.120492 (2021).
50. Dubey, A. K., Kumar, A., Díaz, V. G., Sharma, A. & Kanhaiya, K. Study and analysis of SARIMA and LSTM in forecasting time series data. *Sustainable Energy Technologies and Assessments* **47**, 101474. https://doi.org/10.1016/j.seta.2021.101474 (2021).
51. Zhou, L., Zhao, C., Liu, N., Yao, X. & Cheng, Z. Improved LSTM-based deep learning model for COVID-19 prediction using optimized approach. *Engineering applications of artificial intelligence* **122**, 106157. https://doi.org/10.1016/j.engappai.2023.106157 (2023).
52. Frifra, A. et al. Harnessing LSTM and XGBoost algorithms for storm prediction. *Sci Rep* **14**, 11381. https://doi.org/10.1038/s41598-024-62182-0 (2024).
53. Dixon, S. et al. A Comparison of Infectious Disease Forecasting Methods across Locations, Diseases, and Time. *Pathogens* **11**, 185. https://doi.org/10.3390/pathogens11020185 (2022).
54. Alim, M. et al. Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study. *BMJ open* **10**(12), e039676. https://doi.org/10.1136/bmjopen-2020-039676 (2020).
55. Ma, C. et al. The impact of the COVID-19 pandemic on the incidence and mortality of zoonotic diseases in China. *BMJ Global Health*. **7**, e007109 (2022).

Author contributions

G.S., and J.Y.: Conceptualization, Designed, Funding acquisition. L.L., J.S., and G.S.: Writing - Original Draft, Methodology, Visualization, Writing - Review & Editing. Z.C., and Q.G.: Data collection, and Data curation. All authors commented on previous versions of the manuscript. All authors read and approved the final. L.L., and T.Y. contributed equally to this work.

Funding

This study was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region China (Grant Number:2022D01A311).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-80513-z.

Correspondence and requests for materials should be addressed to J.Y. or G.S.