BMC Infectious Diseases

RESEARCH

Open Access

Evaluating the effectiveness of self-attention mechanism in tuberculosis time series forecasting

Zhihong Lv[1], Rui Sun[1], Xin Liu[1], Shuo Wang[2], Xiaowei Guo[1], Yuan Lv[1], Min Yao[3*] and Junhua Zhou[1*]

Abstract

Background With the increasing impact of tuberculosis on public health, accurately predicting future tuberculosis cases is crucial for optimizing of health resources and medical service allocation. This study applies a self-attention mechanism to predict the number of tuberculosis cases, aiming to evaluate its effectiveness in forecasting.

Methods Monthly tuberculosis case data from Changde City between 2010 and 2021 were used to construct a self-attention model, a long short-term memory (LSTM) model, and an autoregressive integrated moving average (ARIMA) model. The performance of these models was evaluated using three metrics: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

Results The self-attention model outperformed the other models in terms of prediction accuracy. On the test set, the RMSE of the self-attention model was approximately 7.41% lower than that of the LSTM model, MAE was reduced by about 10.99%, and MAPE was reduced by approximately 9.87%. Compared to the ARIMA model, RMSE was reduced by about 28.86%, MAE by about 32.22%, and MAPE by approximately 29.89%.

Conclusion The self-attention model can effectively improve the prediction accuracy of tuberculosis cases, providing guidance for health departments optimizing of health resources and medical service allocation.

Keywords Tuberculosis, Time series forecasting, Self-attention mechanism, ARIMA model, LSTM model

Introduction

Tuberculosis (TB) is a highly infectious disease caused by *Mycobacterium tuberculosis*, primarily transmitted through the air, and bring a serious threat to global public health [1]. According to the "Global Tuberculosis Report 2023" published by the World Health Organization, there were approximately 10.6 million new cases of TB worldwide in 2022, with an incidence rate of 133 per 100,000 people [2]. China is classified as a high-burden country by the World Health Organization, ranking third globally in the estimated number of TB cases in 2022 [3]. Due to the large population base of China, TB brings a severe burden on public health and the economy [4]. Therefore, it is crucial to predict the future number of TB cases to

*Correspondence:
Min Yao
yaomin1984@126.com
Junhua Zhou
zhoujunhua@hunnu.edu.cn
[1]Key Laboratory of Molecular Epidemiology of Hunan Province, School of Medicine, Hunan Normal University, Changsha, Hunan 410013, China
[2]Changsha University of Science and Technology, Changsha, Hunan 410114, China
[3]Hunan Provincial Center for Disease Control and Prevention, Changsha, Hunan 410005, China

optimize health resources and medical service allocation. Achieving these goals requires accurate models for predicting TB cases, serving as an early warning system. Predicting TB cases based on historical data, known as time series forecasting (TSF), is the focus of this study.

Time series forecasting (TSF) methods for diseases can be broadly categorized into statistical-based and machine learning-based approaches. Statistical-based methods, such as the Grey model [5], the exponential smoothing model [6], and the autoregressive integrated moving average (ARIMA) model [7], are widely used due to their effectiveness in capturing the linear characteristics of time series. For example, studies have shown that seasonal ARIMA can be used as a predictive tool for TB cases in China [8]. However, these methods are not suitable for time series with nonlinear characteristics, such as TB case data [9, 10].

To capture the nonlinear features of time series, machine learning methods such as random forests [11], support vector machines [12], and back-propagation neural networks [13, 14] have been introduced into TSF for diseases [15, 16]. Machine learning models offer higher predictive accuracy when dealing with time series that exhibit nonlinear characteristics. For example, a comparative study indicated that the back-propagation neural network model outperformed the ARIMA model in a TSF task for acquired immunodeficiency syndrome [17].

However, traditional machine learning models struggle to capture long-term dependencies [18]. Long-term dependency refers to the reliance on earlier data in the time series for predicting future cases. Early data in the time series often contain critical information, such as seasonal patterns, epidemiological characteristics, and the latency of disease transmission [19]. For example, in China, TB cases typically peak in spring and tend to decline in autumn and winter [20]. If a model cannot effectively utilize the early data in the time series, the accuracy of its predictions may be significantly reduced.

With advances in computer technology, deep learning models such as recurrent neural networks and long short-term memory (LSTM) networks have been successfully applied to time series forecasting (TSF) for TB [21], influenza [22], and other respiratory diseases [23], largely due to their ability to capture long-term dependencies in time series data. Studies have shown that, compared to ARIMA models, LSTM models reduced the prediction error for TB by 12.92–36.79% [24]. However, as the length of the time series increases, LSTM models may "forget" earlier data, leading to a decline in performance [25, 26]. To address this issue, the Google research team proposed a deep learning model based on the self-attention mechanism in 2017. Unlike LSTM, the self-attention network retains early data in the time series, making it suitable for time series of any length. While this approach has been widely applied in fields such as natural language processing [27] and speech recognition [28], its application in TSF tasks for diseases remains limited.

Therefore, this study proposes a model based on the self-attention mechanism to predict TB cases. The core question of this research is: Can the self-attention mechanism effectively improve the accuracy of TB TSF? First, we collected data on TB cases and performed data description and preprocessing. Next, we constructed prediction models for TB cases, including a self-attention model, an LSTM model, and an ARIMA model. Finally, the structures of the three models were optimized and evaluated, with a particular focus on examining the interpretability and generalizability of the self-attention model. Figure 1 presents the main framework of the study.

Methods

Data Collection and Preprocessing

Data Collection and description

In this study, data on TB cases were obtained from the Disease Surveillance Information Reporting Management System of the Changde Center for Disease Control and Prevention. All TB diagnoses adhered to the Diagnostic Criteria for Tuberculosis (WS 288–2017 edition), issued by the National Health and Family Planning Commission of China. The data spans from January 2010 to December 2021, covering the monthly number of TB cases.

This study collected data on TB cases spanning a total of 144 months. The dataset is complete, with no outliers or missing values. We strictly adhered to the temporal sequence when dividing the dataset, ensuring that future data were not accessed prematurely during model training, thereby maintaining the reliability of the evaluation results. Specifically, the training set included data from January 2010 to December 2019, the validation set contained data from January 2020 to December 2020, and the test set used data from January 2021 to December 2021. Table 1 presents the basic statistical description of TB cases for each dataset. The results show that a total of 61,061 TB cases were reported in Changde City from 2010 to 2021, with the highest number of cases in a single month being 654 and the lowest being 176. Figure 2 illustrates the time series of TB cases.

To analyze the trend in TB cases, this study employed the widely used Seasonal and Trend decomposition using Loess (STL) technique to break down the time series into three components: trend, seasonality, and residuals [29]. Figure 3 illustrates the decomposition of the TB time series. The trend component shows an overall downward trend in TB cases in Changde City from 2010 to 2021. The seasonal component reveals that January to February

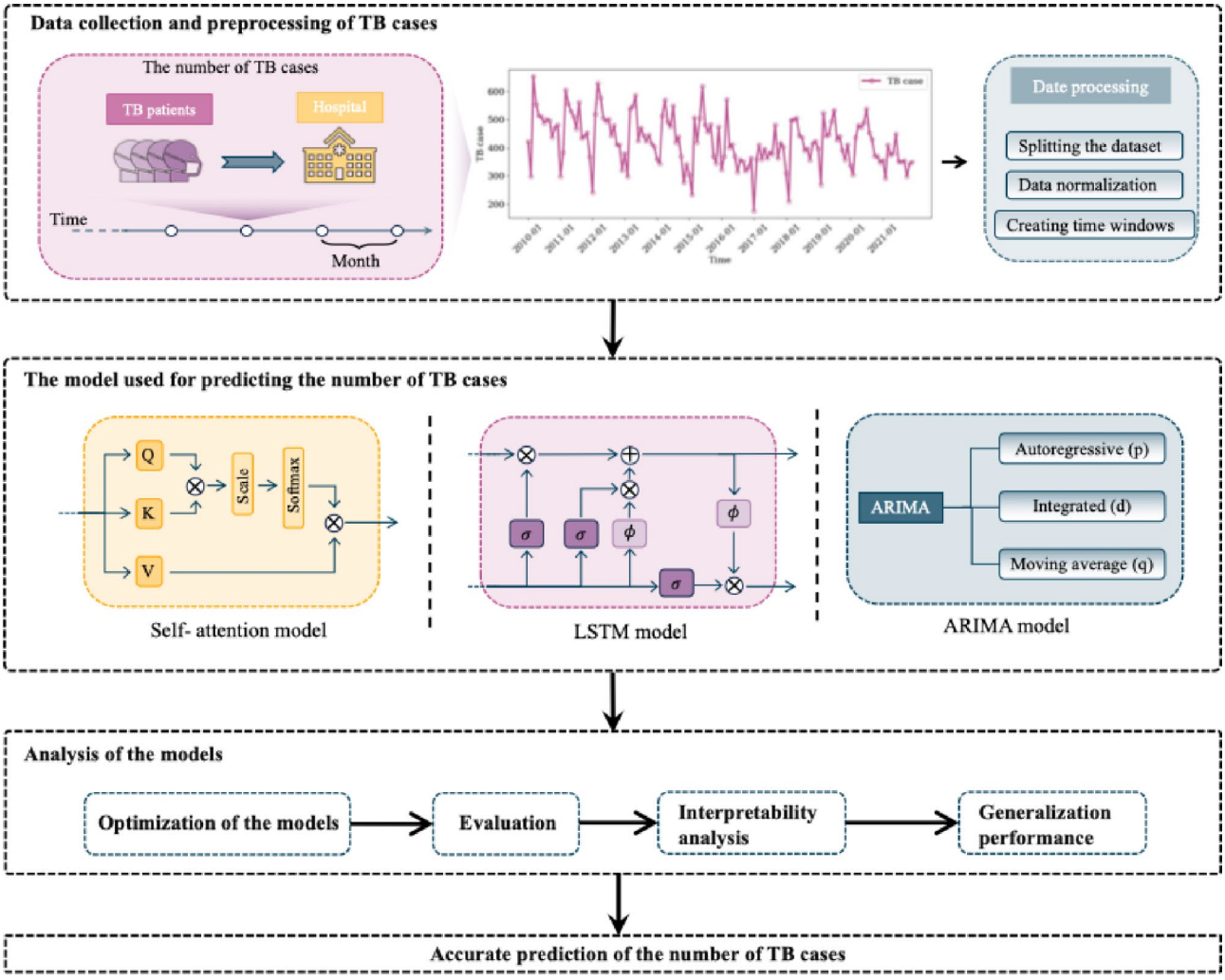

Fig. 1 The main framework of the research in this paper

Table 1 Monthly tuberculosis cases in Changde City (2010–2021): training, testing, and validation dataset description

Dataset	Time range	Total cases	Mean ± SD	Min	P25	Median	P95	Max
Train	2010–2019	51,716	431 ± 90	176	370	427	498	654
Validation	2020	5035	419 ± 72	304	365	434	474	539
Test	2021	4310	359 ± 43	291	347	353	376	450
	2010–2021	61,061	424 ± 88	176	360	421	488	654

is the low period for TB cases, March to June represents the peak period, and July to December marks the declining period.

Data preprocessing

To facilitate faster model convergence, the data needed to be preprocessed [30]. First, the TB cases data were scaled to a range between 0 and 1 using a normalization formula to eliminate differences in variable magnitude. The normalization formula is as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where X_{norm} represents the normalized value, X represents the number of TB cases, X_{min} represents the minimum number of TB cases, and X_{max} represents the maximum number of TB cases.

Next, the normalized data were transformed into input-output pairs using a fixed time window. Figure 4 illustrates the data transformation process. First, the size of the time window was determined. Then, using a sliding window approach with a time step of 1, data were

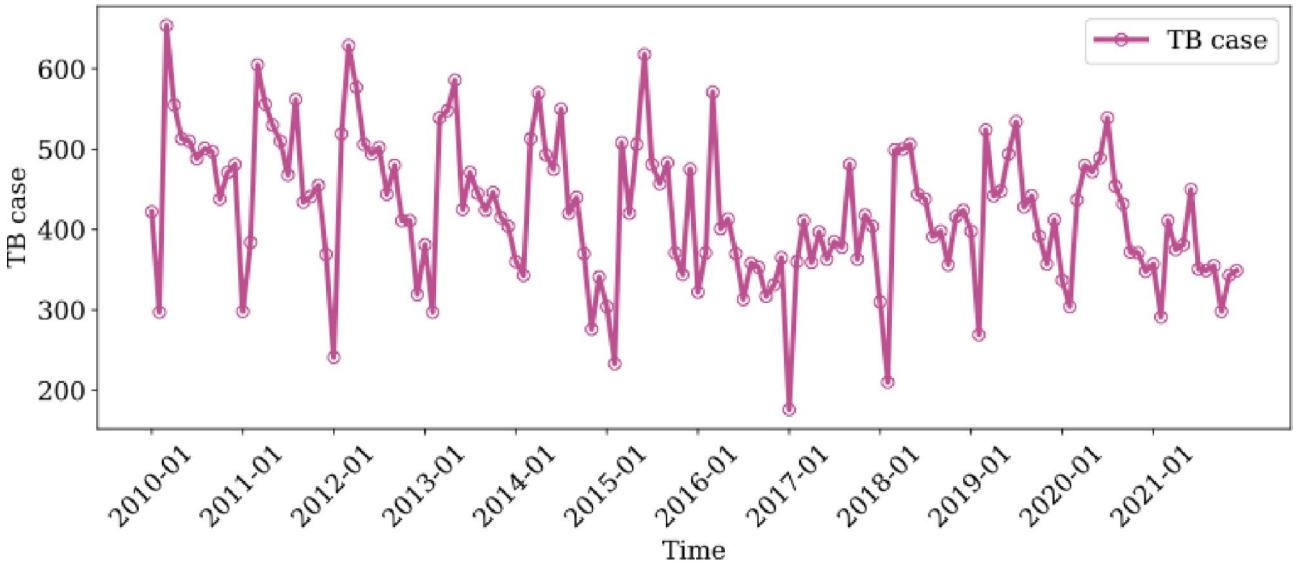

Fig. 2 Monthly time series of tuberculosis cases in Changde City, 2020–2021

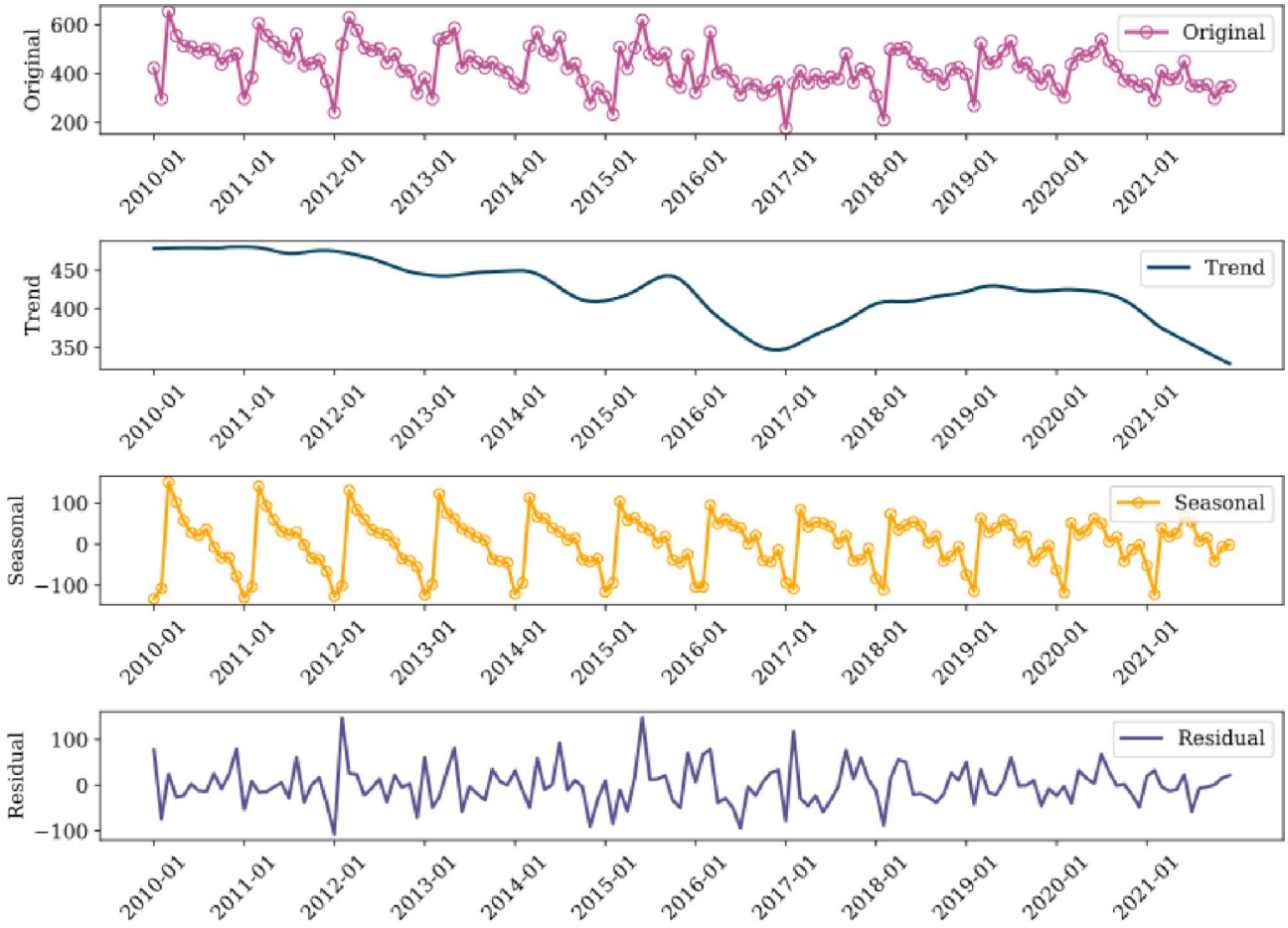

Fig. 3 Decomposition of monthly tuberculosis time series in Changde City, 2020–2021

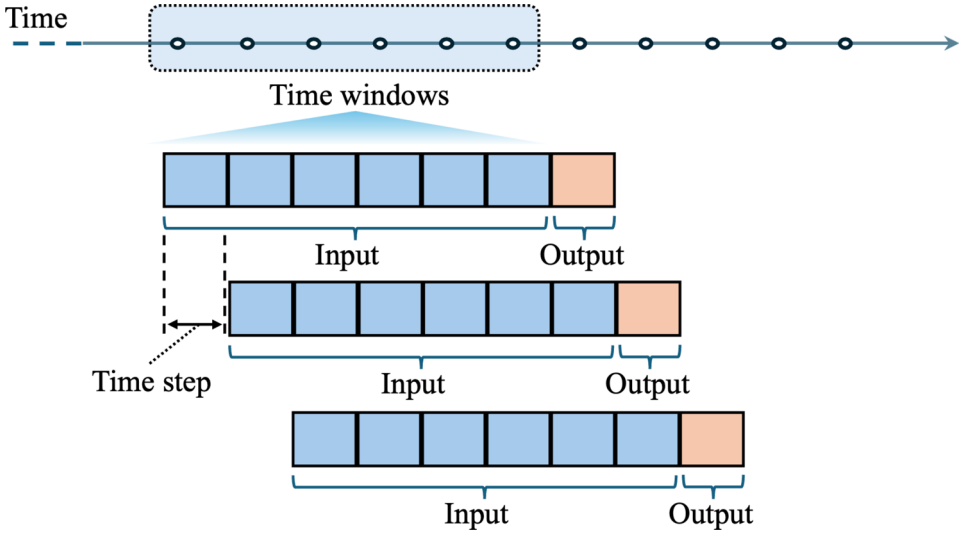

Fig. 4 Flowchart of the data transformation process

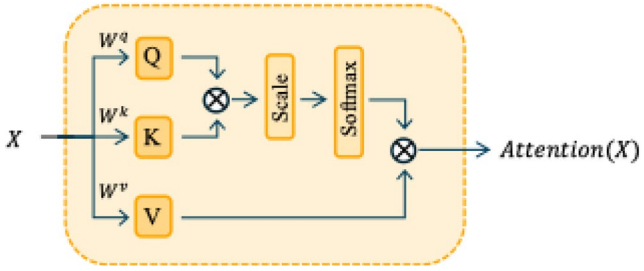

Fig. 5 Schematic diagram of self-attention mechanism

extracted from the TB time series. The data within each time window served as the model's input, while the subsequent data point was used as the model's output. These input-output pairs were then used to train the model, enabling it to predict future TB cases.

Models for Predicting tuberculosis cases

Self-attention model

The self-attention mechanism, which forms a key part of the Transformer model proposed by Ashish Vaswani et al. in 2017, relies on an attention weight matrix to allocate focus within a time series, thereby enhancing the model's ability to capture long-term dependencies [31]. The attention weight matrix reflects the importance of each input data point in generating the output.

In this study, the input data consist of a sequence of historical TB cases over t time windows, denoted as $X = [\chi_1 ? \chi_2 ? \cdots ? \chi_t]$, where χ_t represents the number of TB cases in month t. Figure 5 illustrates the calculation principle of the self-attention mechanism. First, X is mapped through a linear transformation into three vectors: the query vector Q, the key vector K, and the value vector V. Then, the query vector Q is dotted with the transpose of the key vector K to calculate a similarity score, known as the attention weight matrix. Subsequently, the attention weight matrix is multiplied by the value vector V to obtain the self-attention output, denoted as $Attention\,(X)$.

Multi-head self-attention [31] divides the self-attention computation into multiple "heads," with each head responsible for learning different subspace representations of the data. Since multi-head self-attention can analyze input data from multiple perspectives simultaneously, it can capture diverse features and patterns in the data. Therefore, this study employed multi-head self-attention to construct the prediction model. The calculation formula is as follows:

$$Q_i = W_i^q \cdot X$$

$$K_i = W_i^k \cdot X$$

$$V_i = W_i^v \cdot X$$

$$head_i = Attention\,(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

$$MultiHead\,(Q,\,K,\,V) = Concat\,(head_1, \cdots, head_i)\,W^o$$

Where W_i^q, W_i^k, W_i^v, and W^o represent the learnable weight matrices, and i represents the i-th attention head. $head_i$ represents the output of the i-th attention head. The $softmax$ function is used to normalize the similarity scores into a probability distribution. d_k represents the length of the key vector for each attention head and dividing by $\sqrt{d_k}$ is used to maintain numerical stability of the model. $MultiHead\,(Q?K?V)$ represents the output of the multi-head self-attention mechanism.

Figure 6 presents the framework of the self-attention model used in this study. The model consists of three hidden layers: the self-attention layer, the global average pooling layer, and the fully connected layer. The self-attention layer captures the long-term dependency features in the data, the global average pooling layer reduces dimensionality, and the fully connected layer generates the prediction results.

In addition, the model's hyperparameters, which are preset by the developers, determine the structure of the model. We set the ranges of these hyperparameters based on relevant TSF literature and model reports [32, 33]. In this study, the time window was set to 4/6/8/12, the number of attention heads to 2/4/8, the dimension per head to 16/32/64, the number of neurons to 4/8/16/32/64, the batch size to 32/64, the learning rate to 0.001, and the maximum number of epochs to 1000. Since the Adam optimization algorithm combines the advantages of momentum and an adaptive learning rate, it effectively facilitates the model's rapid convergence and stable learning process [34]. Therefore, we used Adam to optimize the model parameters in this study.

LSTM Model

The Long Short-Term Memory (LSTM) network was proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997, specifically designed to address the issues of gradient explosion or vanishing that recurrent neural networks encounter when processing long sequence data [35]. The LSTM network controls the storage, updating, and output of information through the introduction of three gates: the input gate, forget gate, and output gate. The forget gate mainly determines which information should be discarded from the cell state. The input gate decides which new information should be stored in the cell state. The cell state passes information through the entire chain, using the forget gate to remove unimportant data and the input gate to add new information. The output gate is responsible for reviewing the values in the cell state and determining which part of the information should be output.

In this study, the LSTM model consists of two hidden layers: an LSTM layer and a fully connected layer. The hyperparameter settings for the model are as follows: the time window is set to 4/6/8/12, with a time step of 1. The number of neurons is set to 4/8/16/32/64, the batch size is set to 32/64, the learning rate is 0.001, and the maximum number of epochs is set to 1000. The Adam algorithm was used to optimize the model parameters.

ARIMA Model

The Autoregressive Integrated Moving Average (ARIMA) model was proposed by Box and Jenkins in the early 1970s, is primarily used for analyzing and forecasting time series data [36]. The ARIMA model combines three elements: autoregression, integration, and moving

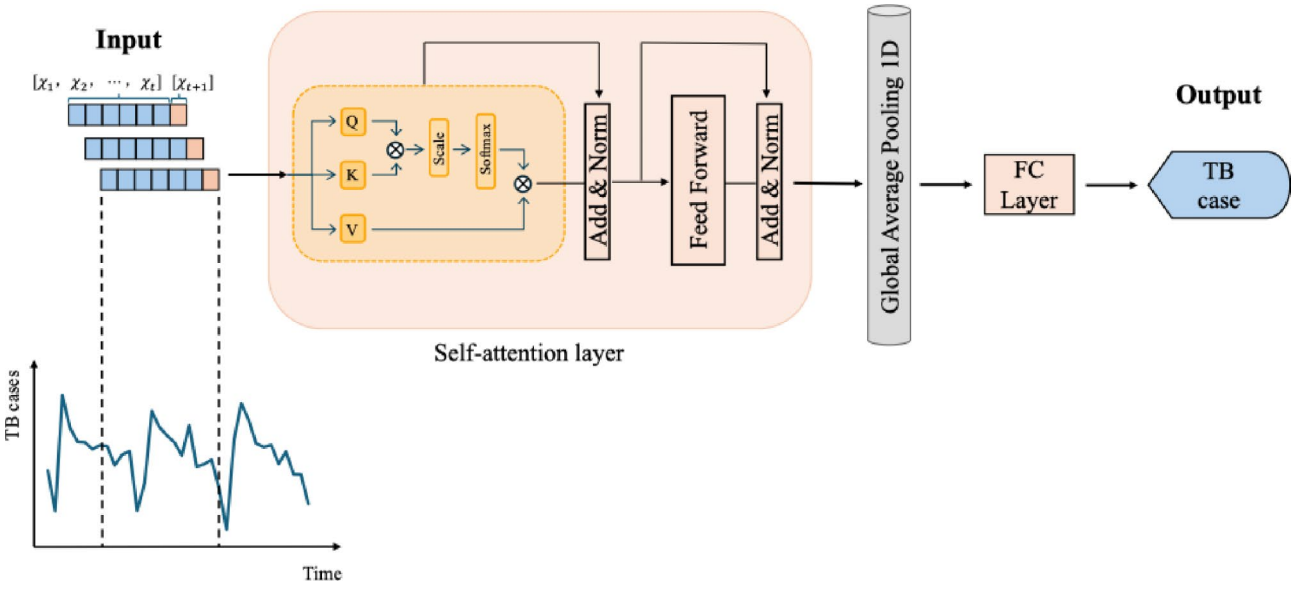

Fig. 6 Structural framework of the self-attention model

average. The model is represented as ARIMA (p, d, q), where p represents the order of the autoregressive terms, indicating the linear relationship between the current observation and its past p observations; d represents the order of differencing needed to make the series stationary; and q represents the order of the moving average terms, indicating the linear relationship between the current error term and its past q error terms.

When the research data exhibit seasonal trends, the seasonal ARIMA model is derived as an extension of the ARIMA model. The seasonal ARIMA model is typically represented as ARIMA (p, d, q) (P, D, Q) S, where P represents the seasonal autoregressive order, D represents the seasonal differencing order, Q represents the seasonal moving average order, and S represents the length of the seasonal cycle. In this study, the establishment of the ARIMA model involved the following steps: First, the time series was visualized, and the Augmented Dickey-Fuller test was used to check if the series was stationary. If the series was not stationary, differencing was applied until stationarity was achieved. Then, the Autocorrelation Function and Partial Autocorrelation Function plots of the stationary series were generated to determine the range of p, q, P, and Q values. Finally, the auto_arima function in Python was used to fit models with different parameters, and the optimal model was selected based on the Akaike Information Criterion, Bayesian Information Criterion, and Ljung-Box Q test.

Model optimization

The self-attention model and LSTM model were optimized using random search for hyperparameter tuning. Compared to grid search, random search is more efficient in larger search spaces, allowing it to find high-performing hyperparameter combinations more quickly [37]. First, multiple hyperparameter combinations were randomly sampled from a predefined hyperparameter space, and the model was trained with each combination. During each training process, the convergence of the model and possible overfitting was assessed by observing the loss curves for both the training and validation sets. Next, the validation set was used to evaluate the performance of each model configuration, with mean squared error used as the evaluation metric. The effects of different hyperparameter combinations were compared, and the optimal configuration was determined based on the combination that performed best on the validation set.

In contrast, ARIMA models have a relatively small parameter space, which makes them suitable for exhaustive exploration using grid search. The optimal parameter combination was selected by thoroughly examining all possible configurations and combining statistical metrics such as the Akaike Information Criterion, Bayesian Information Criterion, and Ljung-Box Q test.

Model evaluation metrics

Currently, root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are widely used to evaluate time series forecasting models [17, 38]. Therefore, this study uses these three metrics to evaluate the performance of each model. RMSE is the square root of the average of the squared deviations between actual values and predicted values. The calculation formula is:

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

MAE is the average of the absolute deviations between the actual values and the predicted values. The calculation formula is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

MAPE is the average proportion of the deviation between actual values and predicted values relative to the actual values. The calculation formula is:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$$

Where n is the number of samples, y_i is the actual TB cases in month i, and $\widehat{y_i}$ is the predicted TB cases in month i.

In this study, RMSE is used to measure the overall deviation of the model when predicting TB cases, MAE evaluates the average deviation, and MAPE measures the relative size of the error, representing the proportion of the error relative to the actual TB cases.

Tools and libraries used

Python 3.10 was used to build the models in this study. TensorFlow 2.15.0 was employed for constructing the self-attention and LSTM models, while the ARIMA model was built using the *statsmodels* 0.14.4 and *pmdarima* 2.0.4 libraries. The statistical significance level was set at 0.05.

Results

Optimal models

Through experimental comparisons, the optimal configuration for the self-attention model was achieved with a time window of 12, 8 attention heads, 32 dimensions per head, a batch size of 64, and 280 training epochs. At this point, the loss curve exhibited a stable decline, reaching its minimum, which indicated the best-performing self-attention model. Similarly, for the LSTM model, the optimal configuration was found with a time window of 12,

16 neurons, a batch size of 64, and 280 training epochs. The loss curve for the LSTM model also showed a stable decline, reaching its minimum, signifying the optimal configuration. For the ARIMA model, the ARIMA (2, 1, 0) (0, 1, 1)12 model had the lowest Akaike Information Criterion and Bayesian Information Criterion values. Additionally, the Ljung-Box Q test ($P=0.92>0.05$) indicated that the residuals were white noise, identifying this as the optimal ARIMA model. Detailed optimization information and residual diagnostics for the three models are provided in Additional file 1 (Supplementary Tables 1 to 4 and Supplementary Fig. 1 to 7).

Table 2 presents a comparison between the predicted and actual values on the test set for the three optimal models. Figure 7 illustrates the fitting and prediction curves of the three models. From the figure, it is evident that during the low TB period in January and February, the self-attention model's predictions were closest to the actual values (with an MAE of 19 cases), followed by the LSTM model (MAE of 30 cases), while the ARIMA model showed a larger deviation (MAE of 39 cases). During the peak period from March to June, both the self-attention model and the LSTM model produced predictions relatively close to the actual values (MAE of 39 cases), whereas the ARIMA model's error increased significantly (MAE of 66 cases). In the declining period from July to December, the self-attention model's predictions remained highly consistent with the actual values (MAE of 23 cases), followed by the ARIMA model (MAE of 25 cases), while the LSTM model showed a slightly larger deviation (MAE of 26 cases).In summary, the self-attention model provided the most stable predictions across all time periods, while the LSTM and ARIMA models demonstrated greater fluctuations in performance.

Table 2 Comparison of predicted and actual monthly tuberculosis cases in Changde City, 2021

Month	Actual	Predicted		
		Self-Attention(AE)	LSTM(AE)	ARIMA(AE)
January	357	325(32)	**331(26)**	291(66)
February	291	**285(6)**	325(34)	302(11)
March	411	**398(13)**	396(15)	500(89)
April	375	431(56)	**368(7)**	446(71)
May	381	424(43)	352(29)	**405(24)**
June	450	**405(45)**	347(103)	371(79)
July	351	425(74)	**375(24)**	408(57)
August	349	340(9)	328(21)	**348(1)**
September	355	**349(6)**	316(39)	364(9)
October	298	336(38)	317(19)	**284(14)**
November	343	**335(8)**	318(25)	299(44)
December	349	**350(1)**	320(29)	326(24)

Note: The comparison is based on predictions from three different models. AE represents the absolute error of the model's prediction

Model performance evaluation

Table 3 presents the performance of the three models on both the training and test sets. The results indicate that, On the training set, the self-attention model achieved an RMSE of 49.72, an MAE of 36.82, and a MAPE of 9.22%, outperforming the LSTM model (RMSE: 57.59, MAE: 44.49, MAPE: 11.28%) and the ARIMA model (RMSE: 62.73, MAE: 47.50, MAPE: 11.96%). On the test set, the self-attention model also demonstrated superior performance, with an RMSE of 35.74, an MAE of 27.54, and a MAPE of 7.58%, compared to the LSTM model (RMSE: 38.60, MAE: 30.94, MAPE: 8.41%) and the ARIMA model (RMSE: 50.23, MAE: 40.63, MAPE: 10.81%).

Additionally, the 95% confidence intervals for the absolute error of the predictions on the test set are [12.42, 42.66] for the self-attention model, [15.62, 46.66] for the LSTM model, and [21.04, 60.23] for the ARIMA model. These results clearly indicate that the self-attention model provides the highest prediction accuracy, followed by the LSTM model, with the ARIMA model exhibiting the lowest accuracy. Figure 8 visually compares the performance of the three models on both the training and test sets, highlighting the superior performance of the self-attention model.

Interpretability analysis of the self-attention model

The interpretability of the self-attention model is primarily achieved by analyzing the attention weight matrix. During each prediction, the self-attention layer generates an attention weight matrix, which represents how TB cases in each month of the input sequence influence TB cases in other months.

Figure 9 illustrates the distribution of the attention weight matrix when predicting TB cases for January and February 2021. In the attention matrix, yellow-colored areas indicate greater attention from the model, while blue-colored areas indicate less attention. Specifically, Figure. 9a shows that when predicting TB cases for January 2021, the model primarily focused on data from January, February, and December 2020. Figure 9b reveals that when predicting TB cases for February 2021, the model placed more emphasis on data from February and December 2020, as well as January 2021. Notably, in both predictions, the model particularly focused on data from February 2020. This suggests that the model considers TB cases from February 2020 to have a significant impact on the predictions for both months.

To further analyze the model's attention allocation during predictions on the test set, we performed a longitudinal summation of the attention matrix to observe how the model focused on different months throughout the entire prediction process. Figure 10 illustrates the model's attention distribution when predicting TB cases for the entire year of 2021. The results indicate that the

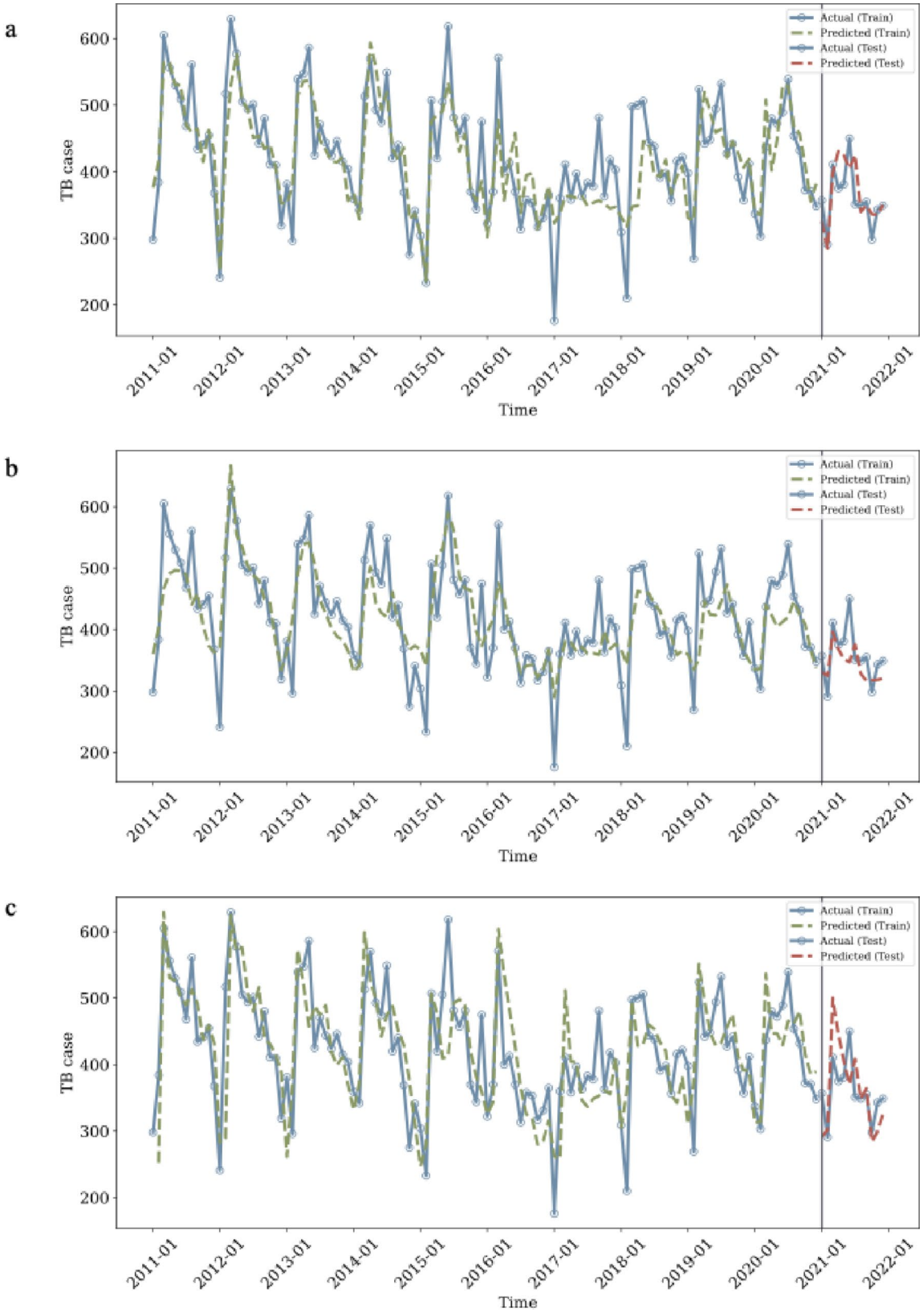

Fig. 7 Fitting and prediction plots of the three models (a: self-attention model, b: LSTM model, c: ARIMA model)

Table 3 Performance comparison of three models on training and test sets

Model	Train set			Test set			
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	AE (95%CI)
Self-Attention	**49.72**	**36.82**	**9.22%**	**35.74**	**27.54**	**7.58%**	**[12.42,42.66]**
LSTM	57.59	44.49	11.28%	38.60	30.94	8.41%	[15.62,46.66]
ARIMA	62.73	47.50	11.96%	50.23	40.63	10.81%	[21.04,60.23]

Note: AE (95% CI) represents the absolute error with the 95% confidence interval

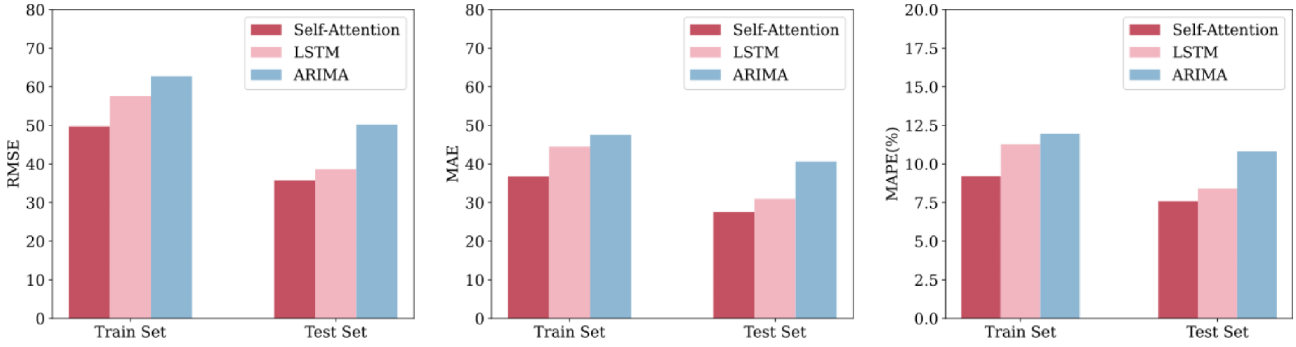

Fig. 8 Performance comparison of the three models

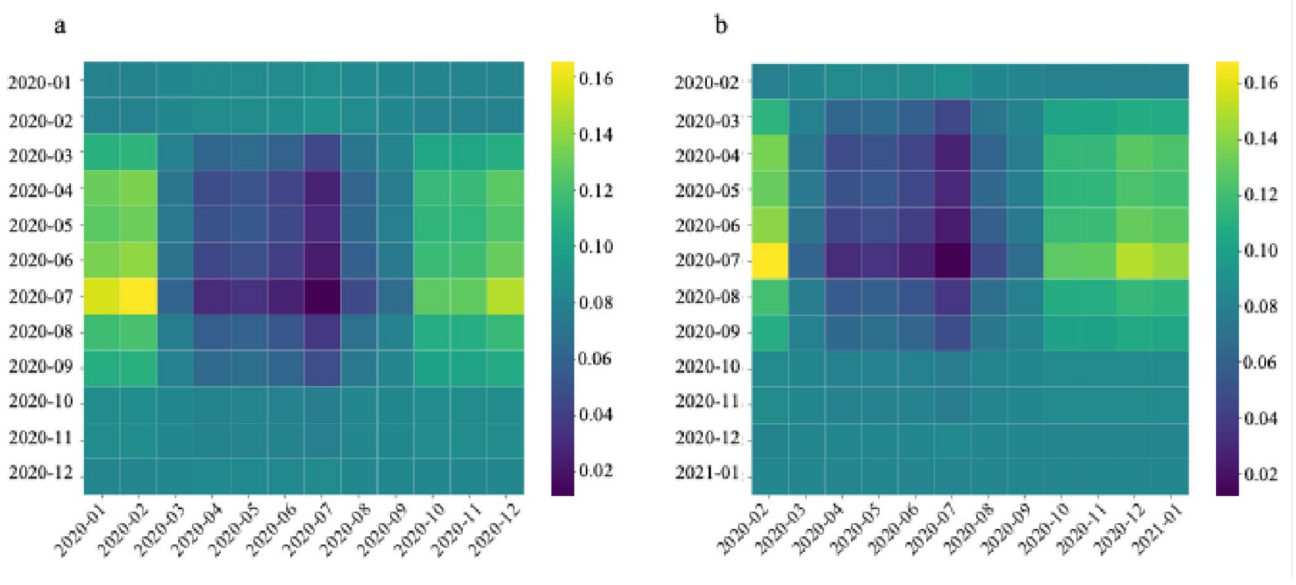

Fig. 9 Attention matrices of the self-attention model for predicting cases: (a) January 2021, (b) February 2021

model particularly focused on data from February 2020 when predicting TB cases for January and February 2021. During predictions from March to October, the model notably focused on data from February 2021. For the predictions in November and December, the model focused primarily on data from February and October 2021. These findings suggest that the model is effective at focusing on early case numbers in the sequence when making future predictions, thereby accurately capturing the long-term dependencies in TB case data. Additionally, Supplementary Figure. 8, included in Additional File 1, presents the model's attention distribution on the training set.

Generalization performance of the self-attention model

To evaluate the self-attention model generalization capability, an external dataset was used for validation. This dataset was sourced from the Chinese Center for Disease Control and Prevention and includes monthly TB cases in Hunan Province from 2004 to 2020. Specifically, data from January 2004 to December 2018 was used as the training set, data from January to December 2019 as the validation set, and data from January to December 2020 as the test set. The self-attention model, LSTM model, and ARIMA model were reconstructed using the same procedures, and their performances were evaluated based on three metrics: RMSE, MAE, and MAPE.

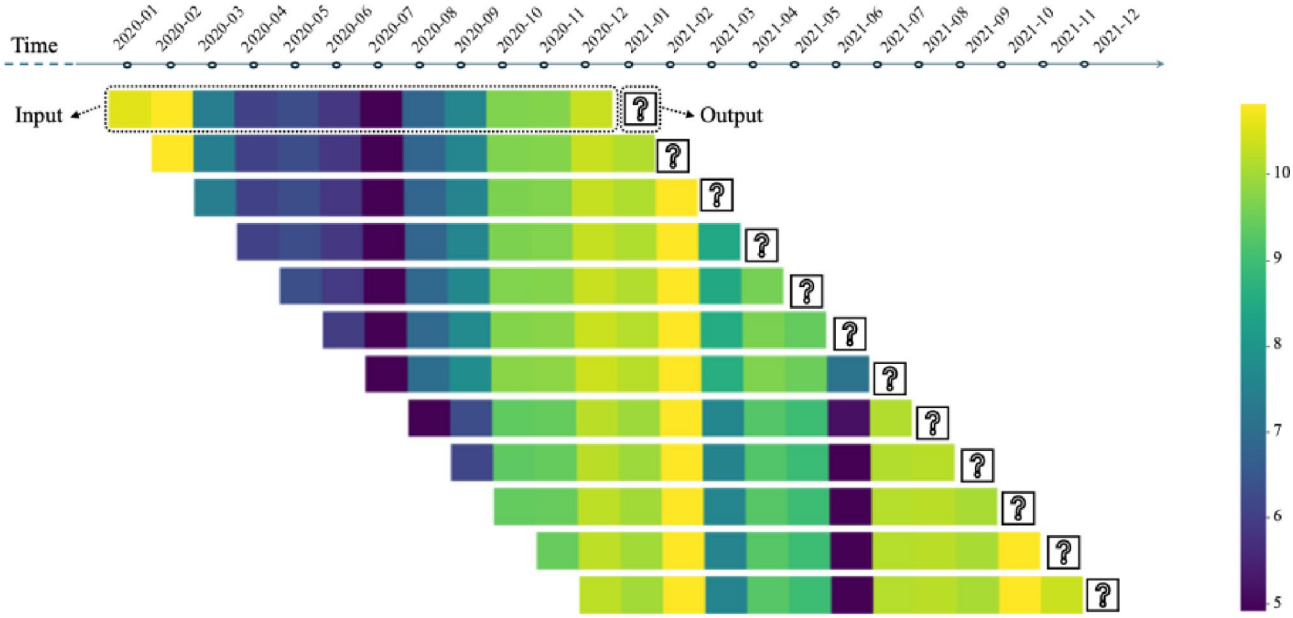

Fig. 10 Attention of the self-attention model in predicting monthly tuberculosis cases in 2021

The model construction and prediction results are presented in Additional File 2. The results show that when predicting TB cases in Hunan Province from January to December 2020, the self-attention model achieved an RMSE of 433.27, an MAE of 274.66, and a MAPE of 6.72%, outperforming the ARIMA model (RMSE: 459.45, MAE: 359.95, MAPE: 8.34%) and the LSTM model (RMSE: 439.23, MAE: 311.06, MAPE: 7.23%). The optimal parameters for the self-attention model were a time window of 12, 8 attention heads, a dimension of 32 per head, a batch size of 64, and 300 epochs. These results indicate that the self-attention model demonstrates strong generalization ability, adapting stably to different regions and time series while maintaining high prediction accuracy.

Discussion

In this study, we applied three time series prediction models to forecast TB cases in Changde City, Hunan Province, China: the self-attention model, the LSTM model, and the ARIMA model. The results indicate that, on the test set, the RMSE of the self-attention model was reduced by approximately 7.41%, the MAE by around 10.99%, and the MAPE by about 9.87% compared to the LSTM model. When compared to the ARIMA model, the RMSE decreased by approximately 28.86%, the MAE by around 32.22%, and the MAPE by about 29.89%.

In terms of specific models, the ARIMA model primarily analyzes the linear characteristics of TB data using statistical methods. However, the TB data also exhibit significant nonlinear characteristics, which the ARIMA model struggles to capture effectively, leading to its suboptimal performance in this study. This finding is consistent with [9], which highlighted that ARIMA performs poorly when dealing with complex nonlinear time series. The LSTM model, on the other hand, has an advantage in handling nonlinear dependencies and retaining useful historical information. However, when dealing with long sequences of data, it may gradually forget earlier information in the sequence [26]. In our predictions of TB cases, the LSTM model did not fully utilize the early data, limiting its prediction performance and reducing its accuracy. In contrast, the self-attention model, through the calculation of attention weights, can capture the potential impact of earlier TB data on later developments, resulting in superior prediction accuracy. This result is consistent with [39], which demonstrated that the Transformer model, based on the self-attention mechanism, outperformed both the ARIMA and LSTM models in predicting influenza incidence.

To understand how the self-attention model predicts TB cases, we conducted an interpretability analysis. Figure 10 illustrates the specific attention patterns of the model across different time periods. When predicting TB cases for various months in 2021, the model primarily referenced data from February 2020, as well as February and October 2021. Notably, these months coincide with traditional Chinese holidays such as the Spring Festival and National Day, during which reported TB cases tend to be lower. Increased social mobility during these holidays may accelerate the spread of *Mycobacterium tuberculosis*, while reduced medical services lead to fewer reported cases. Considering that the average period from TB infection to diagnosis is 72 days (with a standard

deviation of 28 days) [40], this explains the increase in confirmed cases in the months following these holidays [41]. Therefore, the model's focus on data from these months likely aims to capture the delayed effects of holidays on TB incidence and reporting patterns. This demonstrates that the model can effectively allocate attention to time series data, enhancing its ability to capture long-term dependencies.

In addition, we evaluated the generalizability of the self-attention model using an external dataset. The results showed that the self-attention model outperformed both the ARIMA and LSTM models on the external dataset, indicating that it can be effectively used to predict TB case numbers in different regions. While the self-attention model demonstrated strong predictive performance, it does have some limitations. First, the model's high complexity results in long training times and significant resource demands. Second, increased complexity introduces a risk of overfitting, especially when the training data is limited or lacks diversity. Additionally, although this study provided an in-depth analysis of the time series data, it did not account for factors such as climate change, population migration, or improvements in healthcare facilities, all of which may also influence TB incidence.

To address these limitations, future research will focus on reducing the risk of overfitting by increasing the diversity of the dataset. Additionally, future studies will aim to integrate factors like climate and environmental conditions to enhance the self-attention model, creating a more comprehensive and accurate prediction tool.

Conclusion

This study is the first to apply the self-attention mechanism to TB time series forecasting, providing an analysis of both the interpretability and generalization of the model. Compared to ARIMA and LSTM models, the self-attention mechanism can capture the influence of earlier data on future predictions, leading to improved prediction accuracy. This approach can help public health departments implement intervention measures earlier, improving the effectiveness of prevention and control efforts. Additionally, this study opens new possibilities for applying self-attention models to the prediction of other infectious diseases, potentially improving the accuracy of forecasting for other infectious diseases as well.

Abbreviations

TB	Tuberculosis
TSF	Time Series Forecasting
ARIMA	Autoregressive Integrated Moving Average
LSTM	Long Short-Term Memory Networks
STL	Seasonal and Trend decomposition using Loess
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-024-10183-9.

Supplementary Material 1: Table S1. Comparison of Self-Attention Models with Various Parameters. Table S2. Comparison of LSTM Models with Various Parameters. Table S3. Comparison of ARIMA Models with Various Parameters. Table S4. Parameter Test Results for the ARIMA (2,1,0)(0,1,1)12 Model. Figure S1. Training Loss Curves of Self-Attention Models with Various Parameters. Figure S2. Residual Diagnostic Plots for the Optimal Self-Attention Model. Figure S3. Training Loss Curves of LSTM Models with Various Parameters. Figure S4. Residual Diagnostic Plots for the Optimal LSTM Model. Figure S5. Tuberculosis Time Series in Changde City from 2010 to 2020 after First-Order and Seasonal Differencing. Figure S6. ACF Plot (a) and PACF Plot (b) of the Differenced Series. Figure S7. Residual Diagnostic Plots for the Optimal ARIMA Model. Figure S8. Attention of the Self-Attention Model on the Training Set. Additional File 2: Validation of Monthly Tuberculosis Data in Hunan Province. Additional File 3: Code for model building.

Acknowledgements

Not applicable.

Author contributions

ZL, JZ and SW conceived and designed the research. ZL, JZ, RS and XL contributed and checked data. ZL and JZ analyzed the data. ZL wrote the first draft of the manuscript. XG, YL, MY and JZ revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 32370760), the Natural Science Foundation Project of Hunan Province (Grant Nos. 2020JJ5387, 2023JJ20029), the Hunan Provincial Technological Innovation Foundation of China (Grant No. 2023RC3132), and the Scientific Research Program of Hunan Provincial Health Commission (Grant No. C202302088169).

Data availability

Availability of data and materials. All codes and data will be available from the authors upon reasonable request. The model building code is provided in Additional File 3.

Declaration

Ethics approval and consent to participate

This study did not involve any human trials. The use of tuberculosis data from Changde City was approved by the Ethics Committee of the Changde Center for Disease Control and Prevention, Hunan Province, China. The Ethics Review Board of the Changde Center for Disease Control and Prevention deemed that informed consent was not required because the data did not contain personal and health information that could be traced back to the original identifiers. The data used in this study was anonymized before its use. The tuberculosis data from Hunan Province was obtained from the Public Health Science Data Center of the China Disease Prevention and Control Information System (https://www.phsciencedata.cn/Share/), and this data is open to the public.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 24 July 2024 / Accepted: 5 November 2024
Published online: 03 December 2024

References

1. Furin J, Cox H, Pai M, Tuberculosis. Lancet. 2019;393(10181):1642–56. https://doi.org/10.1016/S0140-6736(19)30308-3.
2. WHO. Global tuberculosis report 2023. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023
3. Lv H, Wang L, Zhang X, et al. Further analysis of tuberculosis in eight high-burden countries based on the global burden of Disease Study 2021 data. Infect Dis Poverty. 2024;13(1):70. https://doi.org/10.1186/s40249-024-01247-8.
4. Wang Q, Jiang Q, Yang Y, Pan J. The burden of travel for care and its influencing factors in China: an inpatient-based study of travel time. J Transp Health. 2022;25:101353. https://doi.org/10.1016/j.jth.2022.101353.
5. Yang X, Zou J, Kong D, Jiang G. The analysis of GM (1, 1) grey model to predict the incidence trend of typhoid and paratyphoid fevers in Wuhan City, China. Med (Baltim). 2018;97(34):e11787. https://doi.org/10.1097/MD.0000000000011787.
6. Zhang YQ, Li XX, Li WB, et al. Analysis and predication of Tuberculosis registration rates in Henan Province, China: an exponential smoothing model study. Infect Dis Poverty. 2020;9(1):123. https://doi.org/10.1186/s40249-020-00742-y.
7. Wang YW, Shen ZZ, Jiang Y. Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China. PLoS ONE. 2018;13(9):e0201987. https://doi.org/10.1371/journal.pone.0201987.
8. Mao Q, Zhang K, Yan W, Cheng C. Forecasting the incidence of tuberculosis in China using the seasonal auto-regressive integrated moving average (SARIMA) model. J Infect Public Health. 2018;11(5):707–12. https://doi.org/10.1016/j.jiph.2018.04.009.
9. Saima Rashid YG, Sánchez J, Singh, Khadijah M, Abualnaja. Novel analysis of nonlinear dynamics of a fractional model for tuberculosis disease via the generalized Caputo fractional derivative operator (case study of Nigeria). AIMS Math. 2022;7(6):10096–121. https://doi.org/10.3934/math.2022562.
10. Ruiz-Aguilar JJ, Turias IJ, Jiménez-Come MJ. Hybrid approaches based on SARIMA and artificial neural networks for inspection time series forecasting. Transp Res Part E Logist Transp Rev. 2014;67:1–13. https://doi.org/10.1016/J.TRE.2014.03.009.
11. Nguyen TN, Minh DN. Applying machine learning to predict Hand-Foot-Mouth disease outbreaks in Vietnam. J Health Inf Dev Ctries. 2021;15(2).
12. Lin X, Wang X, Wang Y, et al. Optimized neural network based on genetic algorithm to Construct Hand-Foot-and-Mouth Disease Prediction and early-warning model. Int J Environ Res Public Health. 2021;18(6):2959. https://doi.org/10.3390/ijerph18062959.
13. Ismail S, Fildes R, Ahmad R, Wan Mohamad Ali WN, Omar T. The practicality of Malaysia dengue outbreak forecasting model as an early warning system. Infect Dis Model. 2022;7(3):510–25. https://doi.org/10.1016/j.idm.2022.07.008.
14. Alshanbari HM, Iftikhar H, Khan F, Rind M, Ahmad Z, El-Bagoury AAH. On the implementation of the Artificial Neural Network Approach for Forecasting Different Healthcare events. Diagnostics (Basel). 2023;13(7):1310. https://doi.org/10.3390/diagnostics13071310.
15. Cuba WM, Huaman Alfaro JC, Iftikhar H, López-Gonzales JL. Modeling and analysis of Monkeypox Outbreak using a New Time Series Ensemble technique. Axioms. 2024;13(8):554. https://doi.org/10.3390/axioms13080554.
16. 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. https://doi.org/10.1007/s00284-023-03531-6.
17. 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 Inf Decis Mak. 2020;20(1):143. https://doi.org/10.1186/s12911-020-01157-3.
18. Ubal C, Di-Giorgi G, Contreras-Reyes JE, Salas R. Predicting the Long-Term dependencies in Time Series using recurrent Artificial neural networks. Mach Learn Knowl Extr. 2023;5(4):1340–58. https://doi.org/10.3390/make5040068.
19. Bai, W., & Ameyaw, EK. Global, regional and national trends in tuberculosis incidence and main risk factors: a study using data from 2000 to 2021., https://doi.org/10.1186/s12889-023-17495-6 (2024).
20. Wang Y, Xu C, Ren J, et al. Secular seasonality and Trend forecasting of tuberculosis incidence rate in China using the Advanced Error-Trend-Seasonal Framework. Infect Drug Resist. 2020;13:733–47. https://doi.org/10.2147/IDR.S238225.
21. Li ZQ, Pan HQ, Liu Q, Song H, Wang JM. Comparing the performance of time series models with or without meteorological factors in predicting incident pulmonary tuberculosis in eastern China. Infect Dis Poverty. 2020;9(1):151. https://doi.org/10.1186/s40249-020-00771-7.
22. Zhao Z, Zhai M, Li G, et al. Study on the prediction effect of a combined model of SARIMA and LSTM based on SSA for influenza in Shanxi Province, China. BMC Infect Dis. 2023;23(1):71. https://doi.org/10.1186/s12879-023-08025-1.
23. Tsan YT, Chen DY, Liu PY, Kristiani E, Nguyen KLP, Yang CT. The prediction of influenza-like illness and respiratory Disease using LSTM and ARIMA. Int J Environ Res Public Health. 2022;19(3):1858. https://doi.org/10.3390/ijerph19031858.
24. Yang E, Zhang H, Guo X, Zang Z, Liu Z, Liu Y. A multivariate multi-step LSTM forecasting model for tuberculosis incidence with model explanation in Liaoning Province, China. BMC Infect Dis. 2022;22(1):490. https://doi.org/10.1186/s12879-022-07462-8.
25. Das S, Tariq A, Santos T, Kantareddy SS, Banerjee I. Recurrent neural networks (RNNs): architectures, training tricks, and introduction to Influential Research. In: Colliot O, editor. Machine learning for Brain disorders. Volume 23. New York, NY: Humana; 2023. pp. 117–38.
26. Zhao J, Huang F, Lv J, Duan Y, Qin Z, Li G, Tian G. (2020). Do RNN and LSTM have Long Memory? *International Conference on Machine Learning*.
27. Kumar S, Solanki A. An abstractive text summarization technique using transformer model with self-attention mechanism. Neural Comput Appl. 2023;35(25):18603–22.
28. Subakan C, Ravanelli M, Cornell S, Grondin F, Bronzi M. Exploring self-attention mechanisms for speech separation. IEEE/ACM Trans Audio Speech Lang Process. 2023;31:2169–80.
29. Cleveland RB, Cleveland WS, McRae JE, et al. STL: a seasonal-trend decomposition[J]. J off Stat. 1990;6(1):3–73.
30. Bhanja S, Das A. (2018). Impact of data normalization on deep neural network for time series forecasting. *arXiv preprint arXiv:1812.05519*.
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Adv Neural Inf Process Syst; 2017:5998–6008.
32. Yu T, Zhu H. (2020). Hyper-Parameter Optimization: A Review of Algorithms and Applications. *ArXiv, abs/2003.05689*.
33. Madhusudhanan K, Jawed S, Schmidt-Thieme L. (2024). Hyperparameter tuning MLP's for probabilistic time series forecasting. In D. N. Yang, X. Xie, V. S. Tseng, J. Pei, J. W. Huang, & J. C. W. Lin, editors, *Advances in Knowledge Discovery and Data Mining: PAKDD 2024, Lecture Notes in Computer Science* (Vol. 14650). Springer. https://doi.org/10.1007/978-981-97-2266-2_21
34. Kingma DP. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
35. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
36. Box GEP, Jenkins GM, Reinsel GC, et al. Time Series Analysis: forecasting and control. 5th ed. Hoboken, NJ: Wiley; 2015.
37. Bergstra J. Random Search for Hyper-Parameter optimization. J Mach Learn Res (JMLR). 2012;13(Feb):281–305.
38. Zhang R, Guo Z, Meng Y, et al. Comparison of ARIMA and LSTM in forecasting the incidence of HFMD Combined and Uncombined with Exogenous Meteorological variables in Ningbo, China. Int J Environ Res Public Health. 2021;18(11):6174. https://doi.org/10.3390/ijerph18116174.
39. Wu N, Green B, Ben X et al. Deep transformer models for time series forecasting: the influenza prevalence case[J]. arXiv preprint arXiv:2001.08317, 2020.
40. Guo Z, Xiao D, Wang X, Wang Y, Yan T. Epidemiological characteristics of pulmonary tuberculosis in mainland China from 2004 to 2015: a model-based analysis. BMC Public Health. 2019;19(1):219. https://doi.org/10.1186/s12889-019-6544-4.
41. Zuo Z, Wang M, Cui H, et al. Spatiotemporal characteristics and the epidemiology of tuberculosis in China from 2004 to 2017 by the nationwide surveillance system. BMC Public Health. 2020;20(1):1284. https://doi.org/10.1186/s12889-020-09331-y.

Publisher's note