A Multivariate Multi-Step LSTM Forecasting Model For Tuberculosis Incidence With Model Explanation In Liaoning Province, China

Enbin Yang

Jilin University

Hao Zhang

Jilin University

Xinsheng Guo

Jilin University

Zinan Zang

Jilin University

Zhen Liu

Jilin University

Yuanning Liu (✉ liuyn@jlu.edu.cn)

Jilin University

Research Article

Keywords: Tuberculosis, Machine learning, Multivariate multi-step LSTM model, Hybrid forecasting model, Model explanation, SHAP

Posted Date: January 10th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-986856/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

A multivariate multi-step LSTM forecasting model for tuberculosis incidence with model explanation in Liaoning Province, China

Enbin Yang^{1,2}, Hao Zhang^{1,2,3}, Xinsheng Guo^{1,2}, Zinan Zang^{1,2}, Zhen Liu^{1,4} and Yuanning Liu^{1,2,3*}

*Correspondence: liuyn@jlu.edu.cn

¹College of Computer Science and Technology, Jilin University, 130012, Changchun, China

²Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, 130012, Changchun, China
Full list of author information is available at the end of the article

Abstract

Background: In addition to COVID-19, tuberculosis (TB) is the respiratory infectious disease with the highest incidence in China. We aim to design a series of forecasting models and find the factors that affect the incidence of TB, thereby improving the accuracy of the incidence prediction.

Results: In this paper, we developed a new interpretable prediction system based on the multivariate multi-step Long Short-Term Memory (LSTM) model and SHapley Additive exPlanation (SHAP) method. Moreover, four accuracy measures are introduced into the system: Root Mean Square Error, Mean Absolute Error, Mean Absolute Percentage Error, and symmetric Mean Absolute Percentage Error. Meanwhile, the Autoregressive Integrated Moving Average (ARIMA) model and seasonal ARIMA model are established. The multi-step ARIMA-LSTM model is proposed for the first time to examine the performance of each model in the short, medium, and long term, respectively. Compared with the ARIMA model, each error of the multivariate 2-step LSTM model is reduced by 12.92%, 15.94%, 15.97%, and 14.81% in the short term. The 3-step ARIMA-LSTM model achieved excellent performance, with each error decreased to 15.19%, 33.14%, 36.79%, and 29.76% in the medium and long term. We provide the local and global explanation of the multivariate single-step LSTM model in the field of incidence prediction, pioneering.

Conclusions: The multivariate 2-step LSTM model is suitable for short-term forecasts, and the 3-step ARIMA-LSTM model is appropriate for medium and long-term forecasts. In addition, the prediction effect was better than similar TB incidence forecasting models. The SHAP results indicate that the five most crucial features are maximum temperature, average relative humidity, local financial budget, monthly sunshine percentage, and sunshine hours.

Keywords: Tuberculosis; Machine learning; Multivariate multi-step LSTM model; Hybrid forecasting model; Model explanation; SHAP

Background

Tuberculosis (TB) is a respiratory infectious disease caused by *Mycobacterium tuberculosis* in the lungs. Its existence seriously endangers the safety of people around the world. The World Health Organization (WHO) reports the *Global Tuberculosis Report* in 2019 shows that the global situation of TB is still severe. In 2018, the number of TB cases worldwide reached 70 million, and the number of deaths from TB reached 1.5 million [1]. Shockingly, this number has exceeded the number of HIV/AIDS deaths. The Chinese Center for Disease Control and Prevention

(CDC) classifies TB as a Class B infectious disease (the most serious is Class A). CDC statistics indicate that China has effectively controlled the spread of TB in the past decade. From 2010 to 2020, TB cases declined from 991,350 to 599,587, and the number of deaths from 3,000 to 965 between January to August [2]. However, the incidence of TB in 2019 was 55.55 (1/100,000), far more than respiratory infectious diseases, such as scarlet fever of 5.85 (1/100,000) and whooping cough of 2.15 (1/100,000). Therefore, it is necessary to predict the incidence of TB accurately. The provincial epidemic prevention and control center can evaluate the scale of the epidemic according to the forecast results of incidence rate, and adopt corresponding measures for different areas [3, 4].

The traditional forecasting models are mostly based on one-dimensional time series, including the Autoregressive Integrated Moving Average (ARIMA), grey, Markov forecasting model, etc. Among them, the ARIMA model, as a basic model, is widely used to predict the incidence of various infectious diseases, such as influenza [5], bacillary dysentery [6], tuberculosis [7], AIDS [8], etc., and has been proven it is feasible and effective. In addition, the researchers improved the traditional model based on the characteristics of the disease. When TB has seasonal characteristics, using Seasonal ARIMA (SARIMA) model can achieve better results [7, 9, 10]. In order to include meteorological or air pollution factors in the incidence of TB, the ARIMAX model can be used, and X represents multivariate input [11]. With the development of Machine Learning (ML), more and more studies try to apply ML models to forecast incidence, such as Nonlinear Autoregressive (NAR) neural network [7], General Regression Neural Network (GRNN) [9], Support Vector Machine (SVM) [12], etc. Meanwhile, many studies have proved that the prediction accuracy of ML models is generally higher than traditional models. In order to further improve the prediction performance, the hybrid forecasting model has become a research hotspot [11, 13, 14]. In recent years, in the context of Deep Learning (DL), researchers have been keen to discover the relationship between the incidence of diseases and factors, such as meteorological factors [10, 11], economic and social factors [25], age and professional structure. In addition, it is hoped to obtain the most relevant factor set to improve the prediction accuracy. Feature selection methods include variance selection method [15], Random Forest [16], correlation coefficient method [17], Least Absolute Shrinkage and Selection Operator (LASSO) [18], etc.

Traditional forecasting models have poor adaptability to big fluctuation data. Moreover, there are few studies to explore the performance of models in different periods. Long Short-Term Memory (LSTM) has proven to be a reliable model for forecasting the incidence of infectious diseases [19, 20]. At the same time, its unique forget gate structure may make the model perform well in the long-term forecast. Using a multivariate multistep LSTM model may obtain better prediction results. “Multivariate” can mine the incidence features, and “multistep” can alleviate seasonal and increasing incidence trends [21]. The interpretability of ML models is a research hotspot in the field of ML, and its mainstream methods include Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanation (SHAP) [22]. SHAP has been used to explain Intensive Care Unit (ICU) patient death and the forecasting of nitrogen dioxide concentration [23, 24]. However, in

many ML forecasting models for incidence, no research explained why a particular model achieves an excellent prediction effect.

In this paper, we collected monthly incidence data of TB, 15 sets of meteorological data, and 9 sets of economic and social factors in Liaoning Province from 2005 to 2017, and used LASSO and Random Forest to specify the dimensions of the data set. We established the ARIMA and SARIMA models, and applied the multivariate multistep LSTM model to forecast the incidence of infectious diseases for the first time, and proposed the multistep ARIMA-LSTM hybrid forecasting model. Through the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and symmetric Mean Absolute Percentage Error (sMAPE), the performance of each model is evaluated from short (6-steps ahead), medium (12-steps ahead), and long (24-steps ahead) terms, respectively. Finally, we use the SHAP method to explain the prediction results of the multivariable single-step LSTM model.

Results

Through Eq. 1 and 2, Mean Square Error (MSE) is used for cross-validation, as shown in Fig. 1a. The results show that the optimal λ value in LASSO is 0.0202. On this basis, LASSO eliminated 7 irrelevant factors, thereby reducing the factor set size from 24 to 17. Next, the coefficient compression process is displayed in Fig. 1b. Then, the program of Random Forest was run 5 times, the average value was taken as the importance score of each feature, and the top 10 and top 5 features were selected as the new factor sets, respectively. Thus, we have a set of factors of size 24, 17, 10, and 5.

Fig. 2a indicates that the period from January to July is the high incidence area of TB, while from October to December is the low incidence area. Thus, in general, the incidence of TB in Liaoning Province is a seasonal and periodic bimodal distribution.

The Augmented Dickey-Fuller (ADF) test shows that the p-value of the original time series is 0.50, which is greater than 0.05. Hence the series is not smooth. However, the series after first-order and first-order seasonal differencing reached smoothness with p-values of 1.94×10^{-6} and 0.01, respectively, both less than 0.05, as shown in Fig. 2b and c. Moreover, Fig. 2d indicates that the series becomes

smoother with a p-value of 7.31×10^{-4} after first differencing and then first-order seasonal differencing. Therefore, we set the parameter d of the ARIMA model to 1, and parameter p and q are taken as integers from 0 to 5. Then, model's Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values were calculated. Thus, the 10 models with the smallest AIC and BIC values were first picked, and the five alternative ARIMA models were obtained after taking the common models, as shown in Table 1.

Table 1 A series of alternative the ARIMA $(p, 1, q)$ models (24 steps ahead prediction)

(p, q)	AIC	BIC	RMSE	MAE	MAPE (%)	sMAPE (%)
(5, 5)	408.04	442.55	0.5628	0.4724	11.1506	10.3224
(3, 3)	419.26	442.26	1.3423	1.1488	26.5904	22.5749
(5, 5)	419.61	442.62	0.4672	0.4177	9.9328	9.3198
(5, 5)	427.84	442.21	0.8994	0.7413	16.9383	19.3824
(5, 5)	417.13	445.88	0.9290	0.7617	17.8239	15.7419

Table 2 A series of alternative the SARIMA $(p, 1, q) \times (P, 1, Q)_{12}$ models (24 steps ahead prediction)

(p, q, P, Q)	AIC	BIC	RMSE	MAE	MAPE (%)	sMAPE (%)
(0, 1, 0, 1)	340.64	348.98	0.8471	0.7204	15.9791	14.5258
(1, 1, 0, 1)	341.49	352.60	0.7634	0.6384	14.1518	13.0495
(0, 2, 0, 1)	341.56	352.68	0.7778	0.6519	14.4502	13.2851
(0, 1, 1, 1)	342.54	353.66	0.8323	0.7083	15.7193	14.3164
(1, 2, 0, 1)	342.83	356.73	0.8040	0.6764	14.9867	13.7548

Table 3 Forecast performance of the multivariate LSTM model with different size factor sets

Number of factors	RMSE	MAE	MAPE (%)	sMAPE (%)
24	0.5002	0.3779	9.1007	8.7484
17	0.4499	0.3398	8.3577	7.9290
10	0.4854	0.3502	8.6350	8.0442
5	0.6184	0.4462	11.1905	9.9987
0	0.5213	0.4106	10.2265	9.4457

Notes: The LSTM model with the number of factors of 0 is based on only the prediction of time series.

The test results of the ARIMA (2,1,4) model indicated that the distribution of points in the Quantile-Quantile (QQ) plot (as in Fig. 3) was along 45-degree, as well as the Durbin-Watson (D-W) value was 1.94, which is close to 2, and the p-value was 0.79, which is greater than 0.05, proving that the model applies to this study. Similarly, five alternative SARIMA models are given in Table 2. Generally, it is considered that a MAPE value less than 10% is an accurate forecasting model [14,35]. The error values of the ARIMA and SARIMA models suggest that ARIMA (2,1,4) is the better model with good performance, and its fit and prediction are shown in Fig. 4a. In comparison, the optimal SARIMA model has a MAPE value of 14.1518, which is a poor forecast performance, as presented in Fig. 4b.

The LSTM solver was run 10 times and the average value was taken as the final error result. Table 3 indicates that the data set, filtered by LASSO only, of size 17, is the optimal factor set. Its improvement for model performance is better than the sets of sizes 24, 10, and 5, without adding the step. Meanwhile, it also outperforms the traditional LSTM model based on time series only. Finally, we solved the multivariate multi-step LSTM models with step sizes of 1, 2, 3, and 4, respectively, and observed the best prediction results at 2-step. Compared with the ARIMA model,

the multivariate 2-step LSTM model has better forecasting effects at any stage, as in Fig. 5a, with a reduction in each error ranging from 8.25% to 22.04%, as presented in Table 4.

Similarly, we selected the same range of step sizes as the LSTM model, thus determining the optimal step size for the hybrid model was 3-step. The error results show that the 3-step ARIMA-LSTM hybrid model performs poorer than the ARIMA model in the short-term forecast. Shockingly, this hybrid model demonstrated excellent performance in the medium and long-term forecasts, as in Fig. 5b,

with errors reduced by 13.48% to 36.79%, nearly double compared to the multivariate 2-step LSTM model, as shown in Table 4.

Discussion

This study aimed to design accurate forecasting models for TB incidence to serve as a reference for epidemic prevention and control departments in Liaoning province. At any stage, using a multivariate 2-step LSTM model can reduce the error by 8.25% to 22.04%, and this performance is on the same level as similar studies. In medium and long-term prediction, if pursuing higher accuracy, the 3-step ARIMA-LSTM hybrid model can be used, and each error can be reduced by 13.48% to 36.79%, an effect that is higher than that of similar studies by 10% to 20% [11, 14].

As a basic model, the ARIMA model, is valid for the majority of infectious disease predictions and is therefore widely used as a comparison to evaluate the performance of new models [8, 11, 13]. However, the SARIMA model, was proved to

be an unreasonable model in predicting the monthly incidence of TB in Liaoning province, and in the short term, the error value was even three times higher than that of the ARIMA model, probably due to the less significant seasonal incidence characteristics of the region.

Table 4 Comparison of the forecast performance of each model

Model	RMSE	MAE	MAPE (%)	sMAPE (%)
6-step ahead prediction between January 2016 to June 2016				
A	0.3244 (-)	0.2811 (-)	6.0454 (-)	5.8097 (-)
B	1.0157 (+213.10%)	0.9339 (+232.23%)	20.0035 (+230.89%)	17.9006 (+208.12%)
C*	0.2825 (-12.92%)	0.2363 (-15.94%)	5.0797 (-15.97%)	4.9490 (-14.81%)
D*	0.4659 (+43.62%)	0.3206 (+14.05%)	6.8661 (+13.58%)	7.4156 (+27.64%)
12-step ahead prediction between January 2016 and December 2016				
A	0.4425 (-)	0.3917 (-)	9.7674 (-)	9.1462 (-)
B	0.7825 (+63.40%)	0.6508 (+66.15%)	14.5400 (+48.86%)	13.2301 (+44.65%)
C*	0.4060 (-8.25%)	0.3073 (-21.55%)	7.8076 (-20.06%)	7.5203 (-17.78%)
D*	0.3753 (-15.19%)	0.2619 (-33.14%)	6.1742 (-36.79%)	6.4240 (-29.76%)
24-step ahead prediction between January 2016 and December 2017				
A	0.4672 (-)	0.4177 (-)	9.9328 (-)	9.3198 (-)
B	0.7634 (+63.40%)	0.6384 (+52.84%)	14.1518 (+42.48%)	13.0495 (+40.02%)
C*	0.4108 (-12.07%)	0.3295 (-21.12%)	7.7436 (-22.04%)	7.4895 (-19.64%)
D*	0.4042 (-13.48%)	0.3070 (-26.50%)	6.9664 (-29.86%)	7.2245 (-22.48%)

Notes: The data format $x(y)$, x is the error value and y is the percentage change compared to the ARIMA model. Particularly, (-) indicates the null value. A is the ARIMA model and B is the SARIMA model. The new model proposed in this paper is labeled by superscript *. C* is the multivariate 2-step LSTM model and D* is the 3-step ARIMA-LSTM hybrid forecasting model.

Specifically, there is no fixed incidence trend of monthly incidence of TB in Liaoning Province. For instance, the incidence rate decreased and then increased from January to March from 2013 to 2015, while it showed a linear downward trend from April to July. However, the trends did not follow the previous pattern from January to July between 2016 and 2017. However, the applicability of the SARIMA model for TB cannot be denied, and some such studies have still given favorable conclusions [7, 9, 14].

The factor set size of 17 had the most significant improvement in the prediction of the LSTM model and can be considered to contain the most important features influencing the development of TB. However, the set size of 24 contained excessive irrelevant factors. In other words, there was noise in the data, which led to poor performance. Moreover, the factor sets of sizes 10 and 5 lack important features, and the traditional LSTM model is based only on time series, so they do not improve prediction accuracy.

Meanwhile, the suitable input step size mitigates the increasing trend of incidence. Thus, the forecast performance of the multivariate LSTM model is further improved. However, even in long-term forecasting, the multivariate 2-step LSTM model still achieves better effects than the ARIMA model, perhaps attributed to the unique “forget gate” structure of the LSTM, resulting in excellent memory for previous sequences [36,37].

We make the first attempt to explain ML forecasting models in the field of infectious disease incidence prediction. The multivariate single-step LSTM model is taken as an example, and the SHAP method is applied for model interpretation. A single sample from January 2016 was selected for the local explanation, and the results are shown in Fig. 6. Features in red indicate pushing the predicted value higher, and features in blue mean driving it lower. Each feature pushes the predicted value of the model, from *base value* = 0.25 to $f(x) = 0.21$. All 24 samples of the test set can be expressed, giving a feature impact map, as displayed in Fig. 7. Each line represents a feature, the horizontal axis indicates the number of samples, and the vertical axis denotes the SHAP value of the feature.

Figure 6 Impact of single sample characteristics (January 2016 forecast).

Figure 7 Feature impact (24 samples of the test set).

The 24 samples of the test set, from January 2016 to December 2017, were selected for the global explanation, and the feature density scatter plot was drawn as shown in Fig. 8a. The horizontal axis is the SHAP value, and each row represents one feature. Red and blue indicate high and low feature values, respectively. To be specific, the first feature, *max_temp*, means the maximum temperature, and the

results show that its lower value will instead drive up the predicted value of TB incidence. The fifth feature, *hour_sun*, indicates the monthly sunshine hours, and its higher value pushes up the predicted value. Finally, the scatter of the last five features oscillates around the SHAP value equal to 0, and there is no spread to either side, which suggests that these features are not associated with the predicted values.

The absolute value of SHAP value was first taken and then averaged to serve as the feature importance, as illustrated in Fig. 8b. It can be seen that the five most important features affecting the incidence of tuberculosis in Liaoning Province are maximum temperature (*max_temp*), average relative humidity (*rel_humidity*), local financial budget (*fin_budget*), monthly sunshine percentage (*month_sun*), and monthly sunshine hours (*hour_sun*). This result corresponds to the conclusion of previous research [11, 25].

Based on the results of the SHAP explanation, we can know which are the important features that influence the incidence of TB and what role, positive or negative, these features play in the generation of predicted values by the model.

Our study has some limitations and drawbacks. First, the factors we selected to influence the incidence of TB were all meteorological factors, but did not include air pollution indicators such as PM2.5 and PM10 [10], since China began to document this kind of data only in 2013. Second, the LSTM designed in this paper has only three layers, and a more complex network structure, combined with parameter tuning methods, may obtain better performance [37]. Finally, the interpretability of ML models has gradually become a new research hotspot in ML in recent years.

Conclusions

In conclusion, the prediction accuracy of the multivariate 2-step LSTM model and the 3-step ARIMA-LSTM hybrid model are both better than that of the traditional ARIMA model and LSTM model. In particular, the hybrid models show excellent performance in the medium and long term. Furthermore, the explanation results of the ML forecasting models can lead us to pick more important features that affect the incidence of TB.

Methods

Data sources

Liaoning Province is located in the southern part of northeast China, between 38°43' N and 43°26' N, and between 118°53' E and 125°46' E, with a total area of 148,600 km^2 . From the Public Health Science Data Center (<https://www.phsciencedata.cn/Share/index.jsp>), TB incidence data in Liaoning Province from 2005 to 2017, 156 months, were collected. From the National Meteorological Science Data Center (<https://www.phsciencedata.cn/Share/index.jsp>), 15 sets of corresponding meteorological data were obtained, including temperature, humidity, sunshine duration, precipitation, pressure, etc. From the National Bureau of Statistics (<https://data.stats.gov.cn/easyquery.htm?cn=E0101>), 9 economic and social data were obtained, including Consumer Price Index (CPI), number of tourists, industrial output, etc.

Data cleaning

The data used in this paper has a good consistency and no outliers. The missing values are alternately filled by the k-Nearest Neighbor (kNN) algorithm and the multivariate feature imputation. The principle of kNN is to find the “nearest” numbers of k samples in the data set to target and interpolate with the average value of k samples [26]. Such “nearest” is measured by Euclidean distance. The multivariate feature imputation is an alternative method when kNN fills in all missing values with the same value [27].

Feature selection

We used a combined method, LASSO and Random Forest. The core of LASSO is to compress the coefficient before the irrelevant variable to zero in the regression problem [28]. The regression problem in this paper can be expressed as

$$y_i = \omega^T x_i + b \quad (1)$$

where, y_i , x_i and ω^T represent the monthly incidence of TB, factors, and regression coefficients.

$$J(w) = \frac{1}{m} \sum_{i=1}^m (y_i - \omega^T x_i)^2 + \lambda \sum_{i=1}^m |w_i|, \lambda > 0 \quad (2)$$

In addition, the cost function $J(w)$ is introduced to evaluate the accuracy of the regression model. It is necessary to find the λ that minimizes the value of $J(w)$. Then, Random Forest is used to ranking the feature importance to reduce the dimension of the factor set further.

Division of train set and test set

For all forecasting models, 132 months of data between January 2005 and December 2015 were used as the train set. The test set selected data from January 2016 to December 2017, in which 6, 12, and 24 months after January 2016 were employed as the verification of 6, 12, and 24 steps ahead prediction.

Building ARIMA model and SARIMA model

In the ARIMA model, “AR” represents autoregression, describing the past and present data relationship. “I” denotes differential operational. “MA” stands for moving average, which is the sum of the error terms in “AR” [13, 29]. The ARIMA model is denoted as $ARIMA(p, d, q)$, where p , d and q represent the order of autoregression, difference, and moving average. The SARIMA model is denoted as $SARIMA(p, d, q) \times (P, D, Q)_s$. P , D , and Q represent the order of seasonal model parameters, and s is the cycle length.

ADF test was used to verify the stationarity of the time series. In addition, we used information criteria, AIC and BIC to determine the order of parameters. AIC is an index to evaluate the fitting effect of the model based on entropy, as shown in Eq. 3. BIC strengthened the punishment item and included the sample number n into the assessment, as shown in Eq. 4. Finally, we verify the applicability of these models according to the QQ plot, D-W test, and white noise test.

$$AIC = -2\ln(L) + 2k \quad (3)$$

$$BIC = -2\ln(L) + k\ln(n) \quad (4)$$

where L is the likelihood function and k is the number of model parameters.

A new interpretable prediction system: multivariate multi-step LSTM model and SHAP

In this section, we applied the multivariate multi-step LSTM model to predict incidence in the field of infectious diseases for the first time. And, the model is selected according to four error indicators. Finally, combined with SHAP, a model explanation method, an interpretable prediction system is proposed.

LSTM structure and theory

LSTM is a variant of Recurrent Neural Network (RNN). Compared with RNN, the memory ability of LSTM was significantly improved. LSTM introduces the concept of “gate”, namely forget gate, input gate, and output gate. Forget gate determines how much information about the previous cell is retained. The input of the forget gate at time t includes output h_{t-1} of the hidden layer at time $t-1$ and new input x_t , which is processed by weighting and Sigmoid activation function, as shown in Eq. 5. Input gate controls the input and update of cell state. f_t is the extent to which previous information c_{t-1} is forgotten, and i_t is also the extent to which new information $\tilde{c}_t$ is added to the cell, as shown in Eq. 6 to 8. Then, output gate selectively outputs which parts, as shown in Eq. 9 and 10. The three-layer LSTM structure used in this paper is illustrated in Fig. 9.

$$f_t = \text{sigmoid}(W_f[h_{t-1}, x_t] + b_f) \quad (5)$$

$$i_t = \text{sigmoid}(W_i[h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t]) \quad (7)$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \quad (8)$$

$$o_t = \text{sigmoid}(W_o[h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \tanh(c_t) \quad (10)$$

Definition of “multivariate” and “multi-step”

In the multivariate multi-step LSTM model, if numbers of m factors are selected, “multivariate” means that the input x_t at time t is not just a number, but a row vector of $1 \times m$ dimension. “Multi-step” means that several data before time t should be used to predict the incidence at time t , which is similar to the “sliding window” [21]. The n -step forecasting diagram is depicted in Fig. 10a. The multivariate multi-step LSTM model can fully consider the factors affecting the incidence of TB, and alleviate the increasing trend by setting multi-step, thereby may obtain a higher prediction accuracy.

Model selection

We introduced four error measures to assess the accuracy of these models. The RMSE and MAE were chosen as scale-dependent measures, and the MAPE and sMAPE were selected as scale-independent measures, which are calculated by Eq. 11 to 14.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (11)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (12)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \quad (13)$$

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^n \frac{2 |\hat{y}_i - y_i|}{|\hat{y}_i| + |y_i|} \quad (14)$$

where, $\hat{y}_i$ is the predicted value, y_i is the actual value, and n is the length of time series.

Model explanation

SHapley Additive exPlanation (SHAP) is a post-hoc method of interpretation. SHAP regards all features as “contributors” and constructs an additive model by calculating the marginal contribution of features to the model output, as shown in Eq. 15 [30–32]. And it can explain why ML forecasting models, even some of the “black box” forecasting models, make certain prediction results. Model interpretation can be divided into Local Interpretation and Global Interpretation. Local interpretation explains a single instance, or the relationship between independent and dependent variables in a data subset [33]. Global interpretation is based on the entire data set, or model level interpretation [34].

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{ik}) \quad (15)$$

where, y_i is the predicted value, y_{base} is the base value of the model, and $f(x_{ik})$ is the SHAP value of feature k .

A new method: multi-step ARIMA-LSTM hybrid model

The hybrid forecasting model can combine the strengths of many models. For instance, the ARIMA-LSTM hybrid model takes the predicted value of the optimal ARIMA model and LSTM model as input, uses LSTM for training, and takes the actual value of TB incidence as output, as presented in Fig. 10b. Further, for the first time, we attempted to add multiple steps to the hybrid model, thus removing the increasing trend of the series to improve the prediction accuracy.

Declarations

Acknowledgements

We thank all the editors and reviewers who worked on this article.

Funding

The study is funded by Jilin Province Industrial Innovation Special Funds Project 2019C053-6 and 2019C053-2; National Key R&D Program Project 2020YFB1709800; National Natural Science Foundation of China 61471181; National Key Research and Development Program Project, (National Science and Technology Development Fund) No. 151, 2020. Natural Science Foundation of Jilin Province YDZJ202101ZYTS144. The funder played no role in study design and execution.

Abbreviations

TB: Tuberculosis; ARIMA: Autoregressive Integrated Moving Average; LSTM: Long Short-Term Memory; LASSO: Least Absolute Shrinkage and Selection Operator; kNN: k-Nearest Neighbor; SHAP: SHapley Additive exPlanation; RMSE: Root Mean Square Error; MAE: Mean Absolute Error; MAPE: Mean Absolute Percentage Error; sMAPE: symmetric Mean Absolute Percentage Error.

Availability of data and materials

The data and code are available on GitHub (https://github.com/EnbinYang/tb_prediction_files).

Authors' contributions

EBY and HZ completed the data collection, model building, and model solving. HZ, XSG and ZNZ was responsible for modifying the presentation of the paper. ZL and YNL provided instruction and advice on the structure of the article. All the authors listed have read and approved the manuscript that is enclosed.

Ethics approval and consent to participate

This study was approved by the China Public Health Science Data Center, the Meteorological Science Data Center, and the National Bureau of Statistics. You can use the raw data provided by these sites directly for your study without any administrative permission. The data used in this paper are publicly available statistics, it was not necessary to obtain informed consent.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Author details

¹College of Computer Science and Technology, Jilin University, 130012, Changchun, China. ²Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, 130012, Changchun, China. ³College of Software, Jilin University, 130012, Changchun, China. ⁴Graduate School of Engineering, Nagasaki Institute of Applied Science, 536 Aba-machi, 851-0193, Nagasaki, Japan.

References

1. Harding E: WHO global progress report on tuberculosis elimination. *The Lancet Respiratory Medicine* 2019, 8(1):1.
2. Li T, Du X, Liu X, Li Y, Zhao Y: Implementation Performance of Tuberculosis Control in China: 2011–2020. *China CDC Weekly* 2021, 3(12):252-255.
3. Jeffries C, Lobue P, Chorba T, Metchock B, Kashef I: Role of the Health Department in Tuberculosis Prevention and Control—Legal and Public Health Considerations. *Microbiology spectrum* 2017, 5(2):1-20.
4. Organization WH: Infection Prevention and Control of Epidemic- and Pandemic-Prone Acute Respiratory Infections in Health Care. World Health Organization; 2014.
5. Wang C, Li Y, Feng W, Liu K, Zhang S, Hu F, Jiao S, Lao X, Ni H, Xu G: Epidemiological features and forecast model analysis for the morbidity of influenza in Ningbo, China, 2006–2014. *International Journal of Environmental Research Public Health* 2017, 14(6):559.
6. Li G-Z, Shao F-F, Zhang H, Zou C-P, Li H-H, Jin J: High mean water vapour pressure promotes the transmission of bacillary dysentery. *PLoS one* 2015, 10(5):e0124478.
7. Wang K, Deng C, Li J, Zhang Y, Li X, Wu M: Hybrid methodology for tuberculosis incidence time-series forecasting based on ARIMA and a NAR neural network. *Epidemiology Infection* 2017, 145(6):1118-1129.
8. 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 Medical Informatics Decision Making* 2020, 20(1):1-13.
9. Zhang G, Huang S, Duan Q, Shu W, Hou Y, Zhu S, Miao X, Nie S, Wei S, Guo N: Application of a hybrid model for predicting the incidence of tuberculosis in Hubei, China. *PLoS one* 2013, 8(11):e80969.
10. Ding Z, Li Y, Wang X, Li H, Wang W: The impact of air pollution on the transmission of pulmonary tuberculosis. *Mathematical Biosciences Engineering* 2020, 17(4):4317-4327.
11. Li Z-Q, Pan H-Q, Liu Q, Song H, Wang J-M: Comparing the performance of time series models with or without meteorological factors in predicting incident pulmonary tuberculosis in eastern China. *Infectious diseases of poverty* 2020, 9(1):1-11.
12. Siriwasatien P, Phumee A, Ongkruk P, Jampachaisri K, Kesorn K: Analysis of significant factors for dengue fever incidence prediction. *BMC bioinformatics* 2016, 17(1):1-9.
13. Yuan C, Liu S, Fang Z: Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM(1,1) model. *Energy* 2016, 100(apr.1):384-390.
14. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S: An advanced data-driven hybrid model of SARIMA-NNNAR for tuberculosis incidence time series forecasting in Qinghai Province, China. *Infection drug resistance* 2020, 13:867.
15. Yang C, Zhang W, Zou J, Hu S, Qiu J: Feature selection in decision systems: a mean-variance approach. *Mathematical Problems in Engineering* 2013, 2013:1-8.
16. Zhou Q, Zhou H, Li T: Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. *Knowledge-based systems* 2016, 95:1-11.
17. Ge R, Zhou M, Luo Y, Meng Q, Mai G, Ma D, Wang G, Zhou F: McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC bioinformatics* 2016, 17(1):1-14.
18. Ghosh P, Azam S, Jonkman M, Karim A, Shamrat FJM, Ignatious E, Shultana S, Beeravolu AR, De Boer F: Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. *IEEE Access* 2021, 9:19304-19326.
19. Gu J, Liang L, Song H, Kong Y, Ma R, Hou Y, Zhao J, Liu J, He N, Zhang Y: A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China. *Scientific reports* 2019, 9(1):1-10.
20. Chae S, Kwon S, Lee D: Predicting infectious disease using deep learning and big data. *International journal of environmental research public health* 2018, 15(8):1596.
21. Brownlee J: Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python: Machine Learning Mastery; 2018:123-160.
22. Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares JMR, Bellotti R, Tangaro S: Explainable Deep Learning for Personalized Age Prediction With Brain Morphology. *Frontiers in neuroscience* 2021, 15:1-14.

23. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, Strøm T, Chmura PJ, Heilmann M, Dybdahl L: Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health* 2020, 2(4):e179-e191.
24. García MV, Aznarte JL: Shapley additive explanations for NO2 forecasting. *Ecological Informatics* 2020, 56:101039.
25. Dean HD, Fenton KA: Addressing social determinants of health in the prevention and control of HIV/AIDS, viral hepatitis, sexually transmitted infections, and tuberculosis. *Public Health Reports* 2010, 125(4):1-5.
26. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001, 17(6):520-525.
27. Van Buuren S, Groothuis-Oudshoorn K: mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 2011, 45(1):1-67.
28. Tibshirani R: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 1996, 58(1):267-288.
29. Kırbaş İ, Sözen A, Tuncer AD, Kazancıoğlu FŞ: Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos, Solitons Fractals* 2020, 138:110015.
30. Shapley LS: 17. A value for n-person games: Princeton University Press; 2016.
31. Lundberg SM, Lee S-I: A unified approach to interpreting model predictions. In: *Proceedings of the 31st international conference on neural information processing systems*: 2017. 4768-4777.
32. Ribeiro MT, Singh S, Guestrin C: "Why should i trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*: 2016. 1135-1144.
33. Liang Y, Li S, Yan C, Li M, Jiang C: Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing* 2021, 419:168-182.
34. Díaz G, Coto J, Gómez-Aleixandre J: Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression. *Applied Energy* 2019, 239:610-625.
35. Wang Y, Xu C, Zhang S, Wang Z, Yang L, Zhu Y, Yuan J: Temporal trends analysis of tuberculosis morbidity in mainland China from 1997 to 2025 using a new SARIMA-NARNNX hybrid model. *BMJ open* 2019, 9(7):e024409.
36. Shao X, Kim CS: Multi-step short-term power consumption forecasting using multi-channel LSTM with time location considering customer behavior. *IEEE Access* 2020, 8:125263-125273.
37. Song X, Liu Y, Xue L, Wang J, Zhang J, Wang J, Jiang L, Cheng Z: Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *Journal of Petroleum Science Engineering* 2020, 186:106682.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.docx](#)
- [completedfulldata.xls](#)
- [raweconomicandsocialdata.xls](#)
- [rawmeteorologicaldata.xls](#)