Forecasting Multivariate time-series data using LSTM Neural Network in Mysore district, Karnataka

Stavelin Abhinandithe K[1], Madhu B[2], Somanathan Balasubramanian[3], Sridhar Ramachandran[4]

[1]Assistant Professor, Division of Medical Statistics, School of Life Sciences, [2]Professor, Department of Community Medicine, JSS Medical College [3]Former Director Research, JSS AHER, Mysore, Karnataka, India, [4]HOD, MCA, Ramakrishna Mission Vidyalaya, Coimbatore, India

How to cite this article: Stavelin Abhinandithe K, Madhu B, Somanathan Balasubramanian, Sridhar Ramachandran et al Forecasting Multivariate time-series data using LSTM Neural Network in Mysore district, Karnataka. Volume 13 | Issue 4 | October-December 2022

Abstract

Advanced and precise forecasting of infectious diseases plays a critical role in planning and providing resources effectively. Time series forecasting for non-linear issues are accessible using deep learning techniques. The association between climatic parameters and dengue occurrences was investigated in this work, and a forecasting model was constructed using a deep learning approach called long short-term memory (LSTM). Univariate and multivariate LSTM time series forecast models were developed using meteorological and dengue incidence data from January 2006 to December 2019. For univariate data, the Pycharm/Google Colab platform was implemented, as the deep learning framework Keras, which is one of the models in the machine learning library based on Tensorflow. The Pandas Python package with built-in support for time series data was used for multivariate data. The final model was chosen using the mae loss and the Adam optimizer. Once the model had been fixed, predictions were made using the model. The research showed that the meteorological factors such as maximum temperature at lag 3, minimum temperature at lag 3, maximum vapour pressure (lag 0,1 and 2), minimum vapour pressure at lag 1, and vapour pressure daily mean at lag 0,1,4 are all significant predictors of dengue along with RMSE value of 1.121 . The results indicated that LSTM network has higher prediction accuracy than any other traditional forecasting methods. Timely management of seasonal diseases such as dengue along with meteorological parameters can predict epidemics in the future.

Keywords: *Dengue, LSTM, Prediction, Meteorological variables, RMSE*

Introduction

Infectious diseases which are caused by agents, most probably microorganism impairs human health. Infectious diseases occurring either naturally or purposefully instigated biological threats bear intensifying risk to trigger disease, disability, and death. Infectious diseases are a worldwide threat that transcends political and geographic boundaries such that every individual across the globe is at risk [(1)].International Organization for Migration (IOM) further reports that improved global health can be achieved through evidence from the studies that advise on how to reduce the burden of disease and disability due to infectious diseases in low-middle-income or developing countries, thus eliminating the emerging threats of infectious diseases on global health. Dengue is one of such infectious diseases

Corresponding author:
Stavelin Abhinandithe K,
Assistant Professor, Division of Medical Statistics, Division of Medical statistics, School of Life sciences, JSSAHER, Mysuru-570015.
Email: Stavelin.ak@jssuni.edu.in
Contact No: 8095726333

which is increasing and has doubled over the last years. It is more prone in tropical and subtropical countries [(2)]. For understanding its transmission to humans by a microbe, studying climate change and its effects to the disease is necessary. Temperature, rainfall, humidity, vapour pressure, sunshine are the significant meteorological factors for the spread of the disease. Knowing the relationship between variation in these climatic factors and dengue incidences helps to predict the disease outbreak accurately. Machine learning along with deep learning techniques are used as the powerful predictive techniques to know the influence of climatic factors on dengue incidences [(3)].

Fang et al., (2020) studied the relationship between meteorological elements and air quality by using LSTM neural networks for Beijing data. In his studies, he observes that the LSTM models predicts better as the model with better accuracy and robustness and thus he concludes that LSTM models can be used as prediction methods[(4)]. Chathurangi et al., (2021) describes the machine learning models such as LSTM to analyse dengue incidences along with weather and population density to predict and forecast the dengue incidences [(5)]. In his study, he observes and predicts the incidence of dengue with a good precision level. The recent advances in deep learning have helped the healthcare industry. Various applications have been implemented and commercialized which incorporated AI-driven component that assists doctors and healthcare providers in achieving accurate diagnosis [(6)].

In this study, we have applied LSTM models taking into account the univariate and multivariate time series infectious data along with meteorological parameters. LSTM are distinct type of RNN capable of handling long-term dependences. Machine learning is increasingly being utilized for linear and non-linear time series models to improve predictions. RNN, one of the machine learning models, works with time series to construct network structures, making it more adaptable in time series data analysis. The LSTM model is one of the RNN variations that address the problem of RNN gradient disappearance and explosion, allowing damaged neural networks to be utilized for long-term time series forecasting. LSTM models have turned into a vital model for forecasting time-series and are suitable for situations involving sequences with autocorrelation.

LSTM Architecture

RNN is one of the powerful techniques for deep neural network that plays a significant role in processing long-term dependent time-series data. Hochreiter and Schmidhuber (1970) proposed the problem of gradient descent for long-term dependence is handled utilizing the LSTM neural network model [(7)]. LSTM can encode and decode time series data; hence memory units can be used in place of hidden layer neurons in RNNs to actualize previous knowledge from memory. One or more memory cells, as well as three door controllers, are found in each memory unit.

LSTM have chain like structure. The cell state, which is a kind of conveyor belt that runs straight down the entire chain, is a key structure of LSTM. The ability of LSTM to add or subtract information from the cell state is controlled by a structure known as gates. Gates are the one which allows information through .These gates are made up of a sigmoid neural net layer and a point wise multiplication operation sigmoid[(7)]. layer's outputs are integers between 0 and 1, indicating how much of each component should be processed. A value of zero indicates that 'nothing should be allowed through;' whereas a value of one indicates that 'everything should be allowed through [(8)]**.** As a result, the LSTM contains three gates to protect and control the state of the cell.

The Fundamental building of RNN and LSTM are alike that it has a chain structure. LSTM is not based on single network-layer and four modules interact with each other.

LSTM Time Series Network

In sequence to sequence learning, RNN models are specialists at mapping an input sequence to an output sequence. The length of the input and output does not have to be the same. Two RNNs, such as LSTMs, are used as a sequence to sequence model. These are encoders and decoders, respectively [(9)]. The encoder's job is to convert a given input sequence into a context vector, which is a fixed length vector. To forecast the output sequence, the decoder is given the context vector as input and the final encoder state as a starting decoder state. Speech recognition, language translation, time series forecasting, and other applications use this form of sequence to sequence learning.

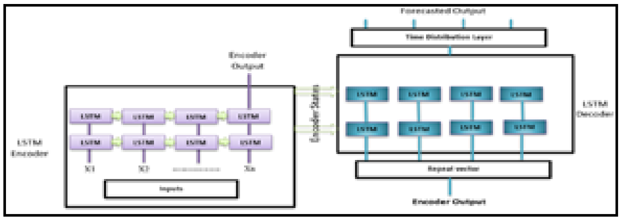


Figure 1: Multivariate Multistep forecasting time series data using stacked LSTM sequence to sequence auto encoder in tensorflow/Keras

Need for applying LSTM

LSTM can automatically handle long and short term dependencies and it can be used when there is a presence of multivariate time series data. In our study we have used LSTM models for uncombined (Without meteorological parameters) and combined (With meteorological parameters) to predict and forecast infectious diseases.

Objectives

The objective of LSTM is to capture temporal dependency in data and preserves back propagated error through time and layers. It is also used to compute new states in the memory cell given old ones.

Data and Methodology

Data

Data of monthly incidence of infectious diseases (dengue) were used from the year 2006-2019. We have used natural logarithm for all the diseases. Since some of the observations are zero which means no case of the infectious diseases are added by 5 and taken natural logarithm (ln). Therefore we have considered logarithm for the dengue incidences.

Data Preprocessing

To fit the data we employed a three-layer stacked LSTM. The LSTM model is separated into three periods for training and prediction: training period (12*5), testing period (12*2), and span period (12*1).

Data Normalization

Machine learning algorithms such as Tensorflow and Keras needs the data to be normalized. Here we have used minmaxscaler to normalize the data

Univariate LSTM (uncombined meteorological variables)

The Environment pycharm/Google colab platform, the deep learning framework are used in Keras which is one of the models in machine learning library based on Tensorflow, Keras are considered to be one of the high-level neural networks which are edited by python. In this scenario, 80% of the data was employed as a training set, with the remaining 20% being used as a testing set.

Multivariate LSTM (combined meteorological variables)

For multivariate data the Pandas library in Python with built-in support for time series data is used. Pandas represented the time series datasets as a series. The read csv() method is the most important function in Pandas for loading CSV data. We have run the algorithm in google colab platform. We have considered combined exogenous variables along with infectious disease data and performed spearman's correlation analysis.

The infectious illnesses dataset must first be prepared for the LSTM. The input variables must be normalized and the dataset must be viewed as a problem of supervised learning. The supervised learning task was posed to forecast infectious diseases based on the previous time step's climatic circumstances. The series to supervised () function was used to modify the dataset. To create a model for LSTM on multivariate input data, we divided the prepared dataset into train and test sets. The input and output variables were then extracted from the training and validation sets. Finally, the inputs (X) are molded into the [samples, time steps, features] 3D structure that LSTMs expect. The LSTM has 50 neurons in the first hidden layer and 1 neuron in the output layer for anticipating infectious diseases. The input form will be a single time step with eight attributes. The Mean Absolute Error (MAE) loss function and the Adam optimizer are used.

The model will be suitable for 50 training epochs with a batch size of 72. Keras resets the internal state of the LSTM at the conclusion of each batch. Finally, the fit() function's validation data parameter is set to the training and test loss during training. At the end of the run, both the training and test losses are shown. After the model has been fitted, the prediction for the complete test dataset is derived, and the forecast is mixed with the test dataset and the scaling is inverted. An error score for the model was produced using forecasts and actual values in their original scale, i.e., the Root Mean Squared Error (RMSE), which provides error in the same units as the variable is calculated.

Cross correlation

It was performed to identify the significant lagged meteorological parameters for the incidences of infectious diseases.

Software's used

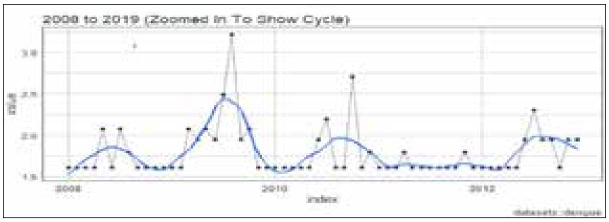
R, Python, keras and tensor flow programming language are used to build LSTM model and stepwise LSTM package was adopted. In this R is used for univariate LSTM with keras and Tensorflow, whereas Python is used for multivariate LSTM with keras and Tensorflow.

Results

Univariate LSTM models for dengue cases

To predict incidence of dengue, we used LSTM model which was developed by using Keras, connects Tensor Flow in R backend. Using the rsample package's rolling forecast origin resampling it performs Time Series Cross Validation with Back Testing. The trend value and forecasted value of dengue for the year 2008-2019 using keras LSTM deep learning is shown in the **Figure 2**.

Figure 2: Fitted and forecasted values of the dengue



for the year 2008-2019

Visualization was done through Sampling Plans and Results Prediction with ggplot2 and cowplot by Keras Stateful LSTM backtested Predictions.

Data is split into train set and test set data in 80% and 20% ratio.

The **Figure 3** shows the data splits into slice 1 and 2.

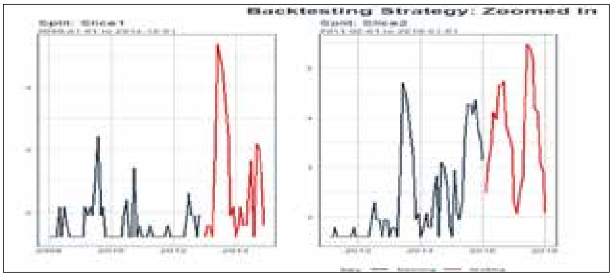


Figure 3: Back testing strategy shows for the year 2008-2014 as slice 1 and for the year 2012-2018 as slice 2

From the **Figure 3**, Slice 1 shows the predicted value for 2014 by using 2008-2012 data and Slice 2 shows the predicted values for 2018 by using 2012-2016.

The Autocorrelation plot was obtained for

Data set of dengue cases

Algorithm of deep learning estimates accurately predicted dengue cases. The backtested Keras Stateful LSTM Model plot was constructed**.**

Table 1: RMSE value based on different slices of data of dengue cases.

Sl. No	Slice No	RMSE	Data
1	1	1.2	2010-2015
2	2	1.3	2011-2019

The RMSE (or another equivalent statistic) average and standard deviation are a helpful approach to compare the performance of different models. The results obtained by using LSTM model are presented in Table 1 .We observe that the mean of the above slices of dengue cases is 1.25 and its standard deviation is 0.05.

From the **Figure 4,** we observe the trend value and forecasted value for dengue cases using Keras LSTM deep learning.

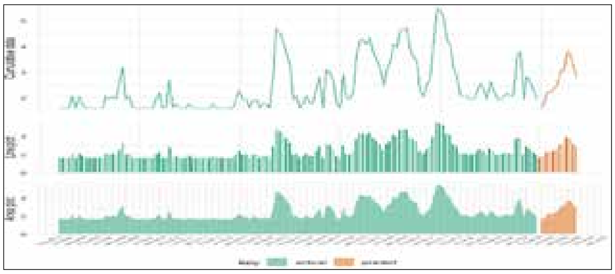


Figure 4: Trend value and forecasted value for dengue cases using Keras LSTM deep learning

We also observed that the Observed and Predicted incidences of dengue that is R^2 value is exactly equal to 1 indicating that all the values are falling on the regression line indicating that the fitted model is best suited for the observed data.

Multivariate LSTM models for dengue cases

To construct multivariate LSTM, we have taken logarithm (ln) dengue data which is taken as dependent variable or targeted variables. Exogenous variables i.e., meteorological variables are considered to be independent variables or features in machine learning language. There are 144 observations in the dengue cases. The whole dataset was split into training and tested data. Correlated exogenous variables are considered along with dengue cases which are entered as columns. Meteorological variables maximum temperature at lag 3, minimum temp_3, maximum vapour pressure (lag 0,1 and 2), minimum vapour pressure at lag 1 and vapour pressure daily mean at lag 0,1,4 are significant predictors for dengue. These series are then converted to supervise learning. Depending on the amount of the data, automatically training and testing data will be considered. Here number of observations is equal to number of hours multiplied by number of features. The next step is to reshape the input to a 3D format that is samples, time steps and features. In order to design the network, epoch will be set 50. Final model will be selected based on mae loss and adam optimizer. Once the model is fixed, prediction will be done based on the model. RMSE value 1.121 was obtained for the dengue.

Discussion and Conclusion

In the current study, a univariate and multivariate LSTM was considered for dengue incidences. The developed model estimates by using existing data of a given month, the number of cases can be estimated with minimal error. Our study has considered the dengue incidence along with 19 meteorological factors to know the impact of climatic change on infectious diseases. We have used machine learning algorithm i.e., Long Short-Term Memory (LSTM) for the infectious diseases for combined with meteorological parameters (multivariate) and uncombined (univariate) data. Based on the results compared with RMSE value for both univariate and multivariate LSTM models, RMSE value for multivariate LSTM is less, indicating that the prediction of accuracy is more when added with exogenous meteorological variables. Further based on Univariate LSTM, we observed that R^2 value is exactly equal to one for dengue incidences indicating that the predicted values fit the observed data exactly and RMSE value for univariate data was 1.25 and for multivariate LSTM was 1.121.Our results indicated that compared to univariate LSTM, multivariate LSTM that is when meteorological variables are added, improves the accuracy of LSTM. The R square value for univariate LSTM was equal to one for dengue.

LSTM models require significantly less time to train in terms of accuracy and computational, and once trained, constant prediction can be estimated. Therefore LSTM models are suitable for predicting the monthly incidence of dengue in Mysuru district, Karnataka.

Ethical clearance - Taken from JSSMC/ IEC/2308/19NCT/2019-20 dated on 3/9/2019 from JSS Medical College. IE Committee.

Source of funding- No funding has been received.

Conflict of Interest - Nil

References

1. De Cock KM, Simone PM, Davison V, Slutsker L. The new global health. Emerging infectious diseases. 2013 Aug; 19(8):1192-7.

2. Miranda JJ, Kinra S, Casas JP, Davey Smith G, Ebrahim S. Non-communicable diseases in low-and middle-income countries: context, determinants and health policy. Tropical Medicine & International Health. 2008 Oct;13(10):1225-34.

3. Fang H, Zhang L. Analysis and Prediction of Air Quality based on LSTM Neural Network——Take Beijing Temple of Heaven as an Example. InJournal of Physics: Conference Series 2020 Sep 1 (Vol. 1637, No. 1, p. 012126). IOP Publishing.

4. Edussuriya C, Deegalla S, Gawarammana I. An accurate mathematical model predicting number of dengue cases in tropics. PLoS Neglected Tropical Diseases. 2021 Nov 8;15(11):e0009756.

5. Said AB, Erradi A, Aly HA, Mohamed A. Predicting COVID-19 cases using bidirectional LSTM on multivariate time series. Environmental Science and Pollution Research. 2021 Oct;28(40):56043-52.

6. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.

7. Khodabakhsh A, Ari I, Bakır M, Alagoz SM. Forecasting multivariate time-series data using LSTM and mini-batches. InThe 7th International Conference on Contemporary Issues in Data Science 2019 Mar 6 (pp. 121-129). Springer, Cham.

8. Wan R, Mei S, Wang J, Liu M, Yang F. Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. Electronics. 2019 Aug;8(8):876.

9. Shah SR, Chadha GS, Schwung A, Ding SX. A Sequence-to-Sequence Approach for Remaining Useful Lifetime Estimation Using Attention-augmented Bidirectional LSTM. Intelligent Systems with Applications. 2021 Jul 1;10:200049.