Weather and population based forecasting of novel COVID-19 using deep learning approaches

A. Ronald Doni¹ · T. Sasi Praba¹ · S. Murugan¹

Received: 3 June 2021 / Revised: 30 July 2021 / Accepted: 12 August 2021 / Published online: 27 August 2021

© The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2021

Abstract The spread of novel corona virus across the globe has a significant impact on various stake holders and posing a major challenge to the research community. Government has taken several measures for maintaining social distance and containment of disease, but still it is not a sufficient for the developing countries like India where the level of understanding the issue is deprived and hence it is a major challenge to the Health Care professionals. Therefore, it is mandatory that a prediction of the number of possible cases enables the preparedness of the Government and the Hospitals in resolving the issues and to take measures in controlling the spread of the disease Series. Deep learning model has been built by considering the features of weather and COVID-19 data (recovered, infected and deceased) for predicting the number of cases expected in India. The model is built on Concurrent Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional RNN (BRNN), Long Short-Term Memory (LSTM) and Bidirectional LSTM (BLSTM) based on the daily weather and COVID-19 data collected from Indian subcontinent. The results revealed that the algorithm BRNN yields a better prediction model when compared with the other models.

Keywords Concurrent neural network (CNN) · Recurrent neural network (RNN) · Bidirectional RNN (BRNN) ·

Long short-term memory (LSTM) and bidirectional LSTM (BLSTM) · COVID-19 · Deep learning

1 Introduction

In December 2019, the first case of COVID-19 has been identified at Wuhan City, China and it turns out to be a major pandemic during the first quarter of the year 2020. The novel corona virus happens to be a major problem in this decade because of its high impact on public health (Togacar et al. 2020). Across the globe, the number of reported cases is 106.61 Crores and the number of deaths is 23.16 Lakhs of which the mostly affected countries are US, India, Brazil, UK and Russia. In India, the number of cases reported during the first week of February 2021 is 1.08 Crores and the number of deceased being 1.55 lakhs (Velásquez and Lara 2020). The virus has its major impact in the elderly people and mostly with multiple health issues and the rate at which the virus spreads is in multiple folds. The virus posts a major challenge to the Government officials, health workers and researchers in controlling the spread and effect of the virus. Various measures like social distance and lockdown across the globe has been implemented for several months to control and prevent the spreading of the virus. It's a great challenge for the researchers to understand the behaviour and the features that have the major influence in spreading or controlling of the virus. Hence, several mathematical models are developed to estimate and predict the number of infected cases and to identify the evolution pattern of the virus (Benvenuto et al. 2020; Ceylan 2020). From the literature it is observed that the models Susceptible, Exposed, Infected and Remove (SEIR) and SIR proved to be the effective

✉ A. Ronald Doni
aronalddoni@gmail.com

¹ Department of Computer Science and Engineering,
Sathyabama Institute of Science and Technology, Chennai,
Tamil Nadu, India

approaches for forecasting the spread of the virus and it is observed that the SIR model is proved to be a better model when compared with SEIR model as per Akaike Information Criteria (Jia et al. 2019; Peng et al. 2020; Roosa et al. 2020; Liu et al. 2020; Zhihua et al. 2020). The models like SIR with Euclidean Network, Generalized Logistic Growth Model, Richards Growth Model and Gompertz Model have also been proposed in predicting the spread of the virus (Biswas et al. 2020; Wu et al. 2020).

In order to facilitate the medical assistance for COVID-19 patients, it is mandatory to predict the number of possible cases for well preparedness and to prevent the loss of live(s). Time-series based prediction of cases is one of the techniques that can be implemented using machine learning and deep learning algorithms. Supervised machine learning algorithms like LASSO regression, Support Vector Machine (SVM) and Exponential Smoothing (ES) have been implemented for predicting the spread of the virus and ES proved to be the best model when compared with the other two approaches (Rustam et al. 2020). When it comes to deep learning approaches LSTM proves to be the best model as it is capable of handling time-based datasets.

In the literature it is evident that deep learning algorithms yields better results when compared with the traditional machine learning algorithms. The survey reveals that the prediction has been carried out for the developed countries and the data set considers the number of cases reported, infected, cured and deceased on day-to-day basis. However, the other parameters like population, health background of the region, climatic conditions, financial viability, education, medical facilities and various other features are not considered. Several studies reveal that the spread of virus has a close association with temperature conditions when tested using epidemiological analysis and mathematical modelling (Lowen et al. 2007; Barreca and Shimshack 2012; Zuk et al. 2009). In the proposed model, weather condition and population features are also included in predicting the COVID-19 cases along with infected, cured and deceased on a daily basis using deep learning algorithms CNN, RNN, BRNN, LSTM, and BLSTM.

The Concurrent Neural Networks (CNN) filters is capable of retrieving the relevant features from the sample input data. The concept of parameter sharing has been implemented in the CNN in which the filter is applied to the various parts of the input in extract the feature map. To address the issue of time dependent learning the concept of Recurrent Neural Networks (RNN) has been developed. The input for the subsequent set of rounds depends on the historical output and the hidden states are maintained. For handling the time series data, the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models are available. Two independent RNN results are integrated and provided as input to the next round. The sequence for

one of the RNN is forward time order and for the other RNNs input is given in the reverse time order. The results of the two networks are concatenated at every iteration and the results are summed up. To process the sequential data, the Bidirectional Long Short-Term Memory (BLSTM) has been introduced. To store the long-range context of information, combination of non-linear and linear feedback loops.

Section 2 discussion about the weather, population and COVID-19 data set, in Sect. 3 the implementation of models using deep learning algorithms and Sect. 4 discusses about the results and performance of the model.

2 Data set

The environmental factors pertaining to a specific region has an inordinate influence in the spread of the disease. In developing countries like India, the factors like population, sanitation, knowledge on hygienic, water, food and climate play a vital role in spreading of virus. The data set related to weather reports (Weather Data Set <https://www.wunderground.com/>; Kaggle 2020; github 2020) and COVID-19 (COVID-19 Data Set: <https://github.com/CSSEGISandData>) is collected from the various sources on day wise starting from January 2020 onwards. The proposed theory aims at identifying the relationship between the weather features like Temperature, humidity, dew, precipitation, wind and pressure across several major regions in India. In a similar manner, number of cases that are infected, deceased and the cases under treatment due to COVID-19 (Dash et al. 2021) are retrieved across all the regions of India on day-to-day basis. Table 1 represents the sample weather data and COVID-19 cases for the city Chennai, Tamil Nadu, India for a period of first ten days in the month of November 2020 and these are the features considered in building the model to predict the spread of the virus. Station wise data on daily basis is collected by the applying the concept of web scrapping. Figure 1 depicts the number of cases reported, deceased and recovered in India for the period from January 2020 to December 2020. In the graph it is observed that the cases are high during the monsoon seasons across India, more specifically in the months of September, October and November 2020. Another major issue to be considered in the rising of cases is due to the release of lock down gradually by the respective State Governments during the period. In India, till the end of August 2020 it is mandatory to register and get approval if the citizen is to move from District to another. However, from September onwards the rule was relaxed, which is also one of the major reasons for the rise in the number of infections.

Table 1 Weather and COVID-19 sample data set

Station	Date	temp_max (°F)	temp_avg (°F)	temp_min (°F)	dew_max (°F)	dew_avg (°F)	dew_min (°F)	humidity_max (%)	humidity_avg (%)	humidity_min (%)	
Chennai	01/11/2020	91	83.3	77	79	76.2	72	100	80.8	55	
	02/11/2020	93	84	77	81	77.1	72	100	80.8	49	
	03/11/2020	90	84.3	77	81	77.8	75	94	81.5	62	
	04/11/2020	90	82	77	81	79.2	77	100	90.8	66	
	05/11/2020	91	83.4	77	84	78.5	75	100	85.5	66	
	06/11/2020	90	82.6	77	81	78	75	100	86.3	66	
	07/11/2020	84	79.8	75	82	77.4	73	100	92.3	79	
	08/11/2020	88	82.2	75	81	74.4	72	100	77.3	66	
	09/11/2020	88	80.9	73	75	69.8	66	94	70.5	49	
	10/11/2020	86	79.8	72	73	69.5	66	94	71.7	51	
Station	Date	wind_max (km/h)	wind_avg (km/h)	wind_min (km/h)	pressure_max (Hg)	pressure_avg (Hg)	pressure_min (Hg)	Percipitation (mm)	Deceased	Hospitalized	Recovered
Chennai	01/11/2020	8	4.3	0	29.8	29.8	29.7	13.0	11	617	792
	02/11/2020	12	4.5	0	29.9	29.8	29.7	13.5	7	662	1238
	03/11/2020	13	4.3	0	29.9	29.8	29.7	11.7	8	653	683
	04/11/2020	14	4.3	0	29.9	29.8	29.7	12.1	5	647	844
	05/11/2020	12	4.9	0	29.9	29.8	29.7	19.5	11	590	681
	06/11/2020	12	4.7	0	29.9	29.9	29.8	20.0	13	600	650
	07/11/2020	10	3.9	0	29.9	29.9	29.8	21.5	5	573	654
	08/11/2020	16	7.7	0	29.9	29.9	29.8	17.0	5	604	675
	09/11/2020	15	7.5	2	29.9	29.8	29.8	8.8	10	557	589
	10/11/2020	10	6.2	2	29.9	29.8	29.7	5.5	8	605	591

Fig. 1 COVID-19 infections in India between January and December 2020

Fig. 2 Impact of climatic conditions and virus outbreak

Figure 2 shows the level of temperature, wind speed and humidity on 22nd July, 2020 in which the number of cases deceased was high.

The observation reveals that the virus spreads extensively when the temperature and humidity is high and it has been observed in the states of Tamil Nadu and Maharashtra.

Apart from the natural factors, the spread of the virus depends on the population on the region of interest. In the

initial days, it is observed that the virus spreads extensively where the population is sufficiently large and the density is high. Figure 3a shows the population of India (projected) as on 30th December 2020 (Suresh et al. xxxx). The source of data set is from the Unique Identification Authority of India (UIDAI) a Government of India organization. Figure 3b shows the population density in India.

To predict the impact of climatic conditions the models are built based on the data set. The size of the data set plays

Fig. 3 **a** Projected population of India (as on 30.12.2020).
b Population density in India

a vital role in the prediction process. The training data set, testing data set and validation data set are randomly chosen. The validation data set is isolated from the model building process (Trappenberg 2019). The formation of model is discussed in the next section.

3 Prediction of COVID-19 cases using weather and population

Figure 4 represents the generic flow of model building using deep learning approaches by consider the COVID-19 data set of India, the weather data that includes temperature, wind speed and humidity and the population data in the Indian subcontinent.

The objective is to identify the correlation between the temperature, wind speed and humidity in spreading the virus. Population is another major attribute in identifying the rate at which the virus spreads. The data pre-processing is the next major task to be computed on the collected data set. The cleaned data is categorized into three sets: training, testing and validation. The percentage of data considered is

in the ratio training: testing: validation is 80:10:10 (Trappenberg 2019). On the training data set the model is built by applying the Deep learning algorithms CNN, RNN and BRNN. Based on the level of accuracy, the model is tuned and the number of epochs is increased accordingly. Finally, the model is tested and validated with the appropriate data sets. The data set reserved for validation is not exposed during the training or testing phase (Lee 2019; Aslam et al. 2021; Bhuyan et al. 2021).

Feature selection is one of the major tasks in data pre-processing. In the proposed work, the features considered are temperature, wind speed, humidity, dew and population to identify the impact of the virus. Random Forest (Paul et al. 2018; Suresh et al. 2021; Homenda and Lesinski 2011) algorithms is applied for identifying the relative importance of the features. Figure 5a, b represents the ranking of features relating to the death and infections due to COVID 19. The feature temperature plays a vital role in the spread of the virus and it is clearly observed in both the number of cases infected and deceased.

In the proposed work, the correlation between the weather attributes and the impact on number of deceased

Fig. 4 Workflow of prediction of COVID-19 by applying deep learning algorithms

and infected COVID 19 cases has been carried out for the Indian Subcontinent. The dependent variable is the number of confirmed COVID-19 cases and is normalized by applying log transformation. The relationship between the temperature, dew, wind speed, rainfall, humidity, population and population density and the COVID-19 case are carried out by applying LASSO regression model. The LASSO regression model has the capability to reduce the impact of the variables that does not have major contribution in the prediction process (Roth 2004). As seen earlier, temperature and humidity have the major impact than the other features, therefore LASSO model has the ability to predict the correlation consequently. Based on the feature selection ranking, the lasso regression is applied on the attribute temperature and humidity. It is observed that, if the recorded average temperature on a given day is less than 80°F then the number of cases registered is less and when the humidity is 70%. Therefore, the threshold for the attribute temperature is set to 80 degrees Fahrenheit and the humidity is set to 70 percentage. The hypothesis is when the temperature and humidity is increased then the rate of spread of virus and number of deaths is also decreased. The

experimental result reveals that there is an inverse relationship between temperature, humidity and the number of infected and deceased cases. The procedure for predicting the number of infected and deceased cases is classified into Model A and B.

Model A (Infected) predicts the number of infected cases against temperature and humidity and Model B (Death) predicts the number of deceased cases against temperature and humidity. Equations 1 and 2 represents the model for predicting the number of possible infections and deaths. The attribute temperature is the independent variable and the dependent variables are humidity and dew. The model A is evaluated based on the total population in the given region and the number of infected cases and model B is computed against the number of deceased cases. The variable α represents the rate of change of temperature on the region of interest and is computed by considering the mean of temperatures recorded. In the similar manner the humidity variable β is the rate of change of humidity and γ represents the rate of change of dew factor recorded in the region. Based on the error rate, the model is adjusted.

a Feature Importance on Covid 19 Infected Cases

b Feature Importance on Covid 19 Deceased Cases

Fig. 5 **a** Feature importance on COVID 19 infected cases. **b** Feature importance on COVID 19 deceased cases

$$A_i : \log(I_c / T_p) = \alpha(t - 80) + \beta(h - 70) + \gamma(d - 75) + \varepsilon \quad (1)$$

$$B_d : \log(D_c / T_p) = \alpha(t - 80) + \beta(h - 70) + \gamma(d - 75) + \varepsilon \quad (2)$$

where A_i —Model A (number of infected cases), B_d —Model B (number of deceased cases), I_c —number of infected cases as on 22nd July 2020, D_c —number of deceased cases as on 22nd July 2020, T_p —total population in the Indian subcontinent, α —the rate of change of temperature, β —the rate of change of humidity, γ —the rate of change of dew factor, ε —training epoch of the neural network.

The model is trained, tested and valuated by applying the deep learning approach Concurrent Neural Network (CNN), Recurrent Neural Network (RNN), Bidirectional RNN (BRNN), Long Short-Term Memory (LSTM) and Bidirectional LSTM (BLSTM) by varying the number of

epochs. The parameters for the deep learning algorithms for high level of accuracy are configured as: learning rate is set to 0.0005, number of hidden layers is 8, epoch is set to 500, timestep is 5. Figure 6a, b shows the model evaluation for deceased and infected cases respectively for the four quarters starting from Jan 2020 to Dec 2020. It is evident that the proposed model predicts the number of infected and deceased cases is almost close to the actuals.

The level of accuracy is 93.23% in case of deceased across all the quarters and for infections it is 92.32%. The results reveals that the temperature, humidity and dew factor play a vital role in the spread of the virus.

4 Results and discussion

The proposed prediction model is evaluated by computing the indexes: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and R-Squared (R^2). Table 1 represents the performance of the model by varying the temperature, humidity and dew factor on both the models and results of the evaluation metrics MSE, RMSE, R-Squared and MAE. To represent the actual differences between the actual and the predicted values in the dataset the metric Mean Absolute Error (MAE) is computed, the variation between the variables is called as Mean Square Error (MSE), the standard deviation is arrived by computing the square root of the MSE and is referred to as Root Mean Squared Error (RMSE). The proportion of variance in the independent variable is represented by R-Squared (R^2) and its value is always less than one (Dash and Dash 2017). Figure 7 represents the evaluation of the deep learning algorithms against the indexes MAE, MSE, RMSE and R^2 .

The study provides a comparison of deep learning algorithms RNN, BRNN, LSTM and BLSTM for forecasting the COVID 19 cases (infected and deceased) in India. By considering the climatic conditions and population in India, the algorithm BRNN provides an enhanced result when compared with the other models. The other features like lock down, health conditions of the infected patients, other climatic conditions may also have a significant impact in the spread of the disease. The impact of the disease after the implementation of vaccination is also to be studied.

5 Conclusion and future scope

The spread of novel COVID-19 leads to the study of impact on climatic conditions and the disease. The factors temperature, humidity, population in a specific region plays a vital role in the spread of the virus. The mathematical

Fig. 6 Prediction of deceased and infected cases using deep learning algorithms

models built on top of the attributes are experimented by applying deep learning algorithms on RNN, CNN and BRNN with LSTM and BLSTM. The factors like complete

lockdown in India from April to May 2020, high temperature due to summer reveals that number of reported cases is low. Once the lock down is lifted and considerably there

Fig. 7 Evaluation of RNN, BRNN, LSTM and BLSTM for predicted COVID 19 infections and deaths

is a reduction in temperature across India, the number of reported cases started increasing. The experimental results reveals that the reduction in temperature leads to the increase in the number of cases. The level of accuracy is high. However, the accuracy can still be increased by regulating the model with a more accurate data set. The proposed model is restricted to the climatic conditions and the population related to the Indian sub-continent only and hence it is necessary to build a generic model which is capable of predicting the spread of the virus. The results suggest the officials to impose lockdown, maintenance of social distancing, medical emergency preparedness and increase the production and consumption of vaccination. At present in European countries, mutant of the novel COVID-19 virus is spreading rigorously, as future work it is proposed to study on the impact on climatic factors in identifying the variant of the virus.

Funding The authors received no specific funding for this study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The manuscript has not been submitted to more than one journal for simultaneous consideration. The manuscript has not been published previously. The Research not involved human participants and/or animals.

References

- <https://github.com/CSSEGISandData/COVID-19>. Accessed 09 Feb 2021
- Aslam B, Javed AR, Chakraborty C, Nebhen J, Raqib S, Rizwan M (2021) Blockchain and ANFIS empowered IoMT application for privacy preserved contact tracing in COVID-19 pandemic. *Pers Ubiquitous Comput*. <https://doi.org/10.1007/s00779-021-01596-3>
- Barreca AI, Shimshack JP (2012) Absolute humidity, temperature, and influenza mortality: 30 years of county-level evidence from the United States. *Am J Epidemiol* 176(suppl7):S114–S122. <https://doi.org/10.1093/aje/kws259>
- Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M (2020) Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data Brief* 105340.
- Bhuyan HK, Chakraborty C, Pani SK, Ravi V (2021) Feature and subfeature selection for classification using correlation coefficient and fuzzy model. *IEEE Trans Eng Manag*. <https://doi.org/10.1109/TEM.2021.3065699>
- Biswas K, Khaleque A, Sen P (2020) COVID-19 spread: reproduction of data and pre-diction using a sir model on Euclidean network. *arXiv:2003.07063*
- Ceylan Z (2020) Estimation of COVID-19 prevalence in Italy, Spain, and France. *Sci Total Environ* 138817.
- Covid-19 Data Set: <https://github.com/CSSEGISandData>. Accessed 22 Feb 2021
- Dash S, Chakraborty C, Giri SK, Pani SK, Frnda J (2021) BIFM: big-data driven intelligent forecasting model for COVID-19. *IEEE Access* 9:97505–97517. <https://doi.org/10.1109/ACCESS.2021.3094658>
- Dash R, Dash PK (2017) MDHS-LPNN: a hybrid FOREX predictor model using a legendre polynomial neural network with a modified differential harmony search technique. In: *Handbook of neural computation*. Academic Press, pp 459–486. ISBN 9780128113189. <https://doi.org/10.1016/B978-0-12-811318-9.00025-9>.
- Homenda W, Lesinski W (2011) Features selection in character recognition with random forest classifier. In: Jędrzejowicz P, Nguyen NT, Hoang K (eds) *Computational collective intelligence. Technologies and applications*. ICCCI 2011. Lecture notes in computer science, vol 6922. Springer, Berlin
- Imdevskp (2020) Covid-19 jhu data web scrap and cleaning. *github*. https://github.com/imdevskp/covid_19_jhu_data_web_scrap_and_cleaning. Accessed 22 Feb 2021
- Jia L, Li K, Jiang Y, Guo X, Zhao T (2019) Prediction and analysis of coronavirus disease. 2020. *arXiv:2003.05447*
- Kaggle (2020) covid19 global weather data. Kaggle. <https://www.kaggle.com/winterpierre91/covid19-global-weather-data>. Accessed 22 Feb 2021
- Lee W-M (2019) Getting started with scikit-learn for machine learning. In: *Python® machine learning*. Wiley, pp 93–117. <https://doi.org/10.1002/9781119557500>
- Liu Z, Magal P, Seydi O, Webb G (2020) Predicting the cumulative number of cases for the COVID-19 epidemic in China from early data. *medRxiv*. <https://doi.org/10.1101/2020.03.11.20034314>
- Liu Z, Magal P, Seydi O, Webb G (2020) Understanding unreported cases in the COVID-19 epidemic outbreak in Wuhan, China, and the importance of major public health interventions. *Biology* 9:50.
- Lowen AC, Mubareka S, Steel J, Palese P (2007) Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog* 3(10):e151. <https://doi.org/10.1371/journal.ppat.0030151>
- Paul A, Mukherjee DP, Das P, Gangopadhyay A, Chintha AR, Kundu S (2018) Improved random forest for classification. *IEEE Trans Image Process* 27(8):4012–4024. <https://doi.org/10.1109/TIP.2018.2834830>
- Peng L, Yang W, Zhang D, Zhuge C, Hong L (2020) Epidemic analysis of COVID-19 in China by dynamical modeling. *arXiv:2002.06563*
- Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman J et al (2020) Real-time fore-casts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. *Infect Dis Modell* 5:256–263
- Roth V (2004) The generalized LASSO. *IEEE Trans Neural Netw* 15(1):16–28. <https://doi.org/10.1109/TNN.2003.809398>
- Rustam F, Reshi AA, Mehmood A, Ullah S, On B, Aslam W et al (2020) COVID-19 future forecasting using supervised machine learning models. *IEEE Access*.
- Suresh P, Saravanakumar U, Iwendi C, Mohan S, Srivastav G (2021) Field-programmable gate arrays with low power vision system using dynamic switching. *Comput Electr Eng* 90:106996
- Suresh P, Aanandhasaravanan K, Iwendi C, Ibeke E, Srivastava G (2021) An artificial intelligence-based quorum system for the improvement of the lifespan of sensor networks. *IEEE Sensors J* 1–13
- Togacar M, Ergen B, Comert Z (2020) COVID-19 detection using deep learning models to exploit social mimic optimization and structured chest x-ray images using fuzzy color and stacking approaches. *Comput Biol Med* 103805.

- Trappenberg TP (2019) Machine learning with sklearn. In: Fundamentals of machine learning. Oxford University Press, pp 38–65. <https://doi.org/10.1093/oso/9780198828044.003.0003>
- Velásquez RMA, Lara JVM (2020) Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. *Chaos Solitons Fractals* 109924.
- Weather Data Set: <https://www.wunderground.com/>. Accessed 22 Feb 2021
- Wu K, Darcet D, Wang Q, Sornette D (2020) Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world. [arXiv:2003.05681](https://arxiv.org/abs/2003.05681)
- Zuk T, Rakowski F, Radomski JP (2009) Probabilistic model of influenza virus transmissibility at various temperature and humidity conditions. *Comput Biol Chem* 33(4):339–343. <https://doi.org/10.1016/j.compbiolchem.2009.07.005>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.