ISSN: 0974-8571 Vol.13 No. 2 December, 2021

International Journal of Computational Intelligence in Control

Optimized Deep Learning Based Ensemble Model for Forecasting of Covid-19

Dr. M. LALLI, Assistant Professor in Computer Science, Bharathidasan University, Tiruchirappalli, Tamilnadu.
M. PANDIJOTHI, Research Scholar in Computer Science, Bharathidasan University, Tiruchirappalli, Tamilnadu.

Date of Submission:10th August 2021 Revised: 25th October 2021 Accepted: 13th December 2021

Abstract; Corona Virus is spreading throughout the world very quickly. COVID-19 is a breathing infection that shows typical signs including fever, shortness of breath, cough, trouble breathing and respiratory symptoms, fever. In more cases, influenza, acute serious respiratory disease, renal failure and mortality may be caused by infection. More than 400,000 people worldwide have been infected by the end of the day on 24 March 2020. For certain vulnerable populations, especially aged, faint-hearted or with multiple chronic conditions, the risk of serious effects from COVID-19 is greater. While the situation in Americas and Europe is getting worse after China has become prosperous, the situation in South Asia is steadily worsening. Through this research work, Deep Learning techniques and Optimization method are utilized to build a ensemble regression model for the forecasting of Covid-19. The deep learning techniques like Deep Auto Encoder, Recurrent Neural Network, Long-Short Term Memory and Genetic Algorithm optimization technique is utilized in this work.
KEYWORDS: Covid-19, Deep Learning, Deep Auto Encoder, Recurrent Neural Network, Long-Short Term Memory, Genetic Algorithm

1. INTRODUCTION

India is the second largest populated region in the world after China in the South Asian nation. The effect of nearly one sixth of the world's population may be attributed to the uncontrolled pandemic of India. In terms of business, agriculture, defence, culture, entertainment, manufacturing and services, India holds important importance on the world map. Indians are present in virtually every part of the world and work for the improvement of world economy and working processes. Trends and estimates for the Indian region must be carefully evaluated to establish successful regional strategies. As the mechanisms for the transmission of COVID-19 are not well known, the numbers of infected individuals are high and the containment effects are largely empirically tested. Thus, it may be important to examine the growth of the outbreak. While COVID-19 spreads worldwide, it also forces economies of the nation's[3], in addition to health risks. The pandemic COVID-19 has impacted populations and economies all over the world in unprecedented ways.

The countries are early in taking a number of diseases stopping steps and their established medical capacity is inadequate to treat patients. While countries follow different disease prevention strategies, they have begun to take collective action to prevent their spread by following each other's strategies as soon as possible.

From an empirical point of view several study models such as ARIMA and Exponential smoothing have been performed in time-series. These are standard techniques that provide a reasonable forecast of time series data in a short time frame. Their broad recognition in the scientific community and fast implementation for different stakeholders have been used to pick the techniques. Deep learning and machine learning are the leading elements for the successful forecasting of time series data instead of the predictive time series model.

2. RELATED WORKS

Feroze, Navid [4] In the Timing Window of March 1st 2020 to June 29th 2020 the Bayesian Structural Time Series (BSTS) models were used to explore COVID-19 temporal dynamics in the top five countries in the world. The authors also assessed, using intervention analyses in accordance with BSTS models, the casual effect of the lockdown on such countries.

Abu-Rayash, Azzam and Ibrahim Dincer [5] The effect of COVID-19 on transport, with corresponding effects on sector-specific power consumption and greenhouse emissions, was

analysed. Including transport quality, technology convergence, congestion rate and accessibility ratio, the smart transport model is suggested. While previous health crises such as SARS have affected transport, the COVID-19 pandemic has been unprecedented and has had a remarkable effect on that sector.

Tandon, Hiteshi, et al [6] The model was developed and then used to predict potential cases in India of COVID-19. The analysis suggests an upward trend in the next few days for the cases. There is also an exponential growth in the number of cases in a time series analysis. The new prediction model is expected to help the government and medical professionals plan themselves for potential problems and make them more equipped for healthcare systems.

Chandu, Viswa Chaitanya [7] Noted that the study aims to take comparative account of in the near future the development of COVID 19 between these two countries, the apparent differences of transmission among the two South East Asian countries of Thailand and India.

Roy, Santanu, Gouri Sankar Bhunia, and Pravat Kumar Shit [8] Spatial distribution in the GIS model was carried out for the study of disease risk by weighted overlay. With the Autoregressive Integrated moving Average (ARIMA), epidemiological trends of prevalence and incidence of COVID-2019 are anticipated.

Mele, Marco, and Cosimo Magazzino [9] The connection among pollution emissions, economic development and deaths in India has been investigated using two separate approaches. Stationarity and Toda-Yamamoto causal studies have been carried out using a time series method and annual data for the years 1980 to 2018.

Ouldali, Naim, et al [10] In the past 15 years, the French highest point of the COVID-19 epidemic carried out a quasi-experimental disrupted time series study in a Tertiary Pediatric Center in the Paris area. The key result is an approximate quasi poisson regression, the number of Kawasaki cases over time.

Salgotra, Rohit, Mostafa Gandomi, and Amir H. Gandomi [11] For confirmed cases (CCs) and cases of death (DCs) in three main countries, most affected namely Delhi, Gujarat, and Maharashtra as well as throughout India, prediction model models based on genetic programmes (GPs) have been created. The proposed prediction models are described in an explicit form and the prediction variables are analysed in their impotence. Statistical criteria and metrics for measuring and validating advanced models have been used here.

Guo, Tao, et al [12] In some nations and continents, there are increased numbers of confirmed cases and death rates of 2019 corona viral disease (COVID-19). There is no detail about the consequences on the fatal outcomes of cardiovascular injury. In the patients with and without elevation of troponin T (TnT) levels, population data, laboratory results, comorbidities and treatments have been collected and analysed.

Vokó, Zoltán, and János György Pitter [13] aimed to define the flow change point in every European country for the COVID-19 epidemic and assess the correlation of the social distancing level with the observed decline in national epidemics. Time series studies in 28 European countries have been disrupted. The Google Group Mobility Studies have measured the social distance index. Threshold regression estimates of transition, national results evaluated regression by Poisson and social distancing effects in mixed effects of the Poisson regression model.

3. RECURRENT NEURAL NETWORK MODEL

RNN (Sequence Models) are notable techniques in the DL neural network model [14]. RNN is utilized to perceive created picture and text message with greater performance. RNN find difficult to catch the long term relevancy that associates the task based of consecutive connections and the vanishing gradient. With respect to the variant sequence model, the long short-term (LSTM) is foreseen towards catching the long-term associations. The essential objective of Long Short-Term Memory technique intended for get done with the vanishing gradient utilizing the optimizing algorithm to find out the weights of neural systems to prevent the long-term issues of dependency. This RNN (sequence model) is used in this work to forecast the covid-19.

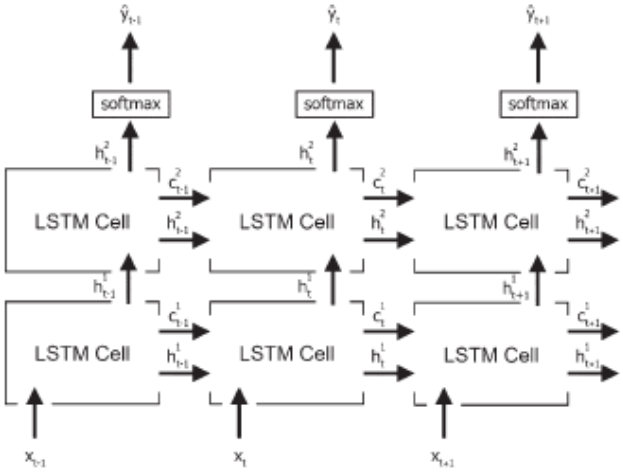

Figure 1: long short-term memory (LSTM) model stacked one another

4. PROPOSED ENSEMBLE REGRESSION MODEL FOR FORECASTING OF COVID-19

4.1 Architecture of Deep Auto-Encoder

Deep Auto-encoder (DA) [15] falls under the category of deep learning technique comprised

of multilayer network system of feed forward type by similar amount of data with respect to input neuron and output neurons. The objective of the DA is to result in reduced representation also decrease error in the data. Model training is done utilizing the back propagation technique as per the loss computation.

Auto encoder approach with more hidden layers is known as the DA (Deep Auto encoder). As numerous encoder and decoder layers are available, it empowers a DA to outline the complex data allocation. Figure 1 depicts the design structure of Auto-encoder consists of first layer as input, two consecutive layer as intermediate and one final layer as output.

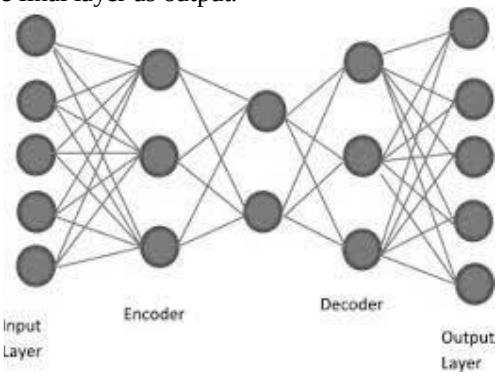

Figure 2: Deep Auto-encoder structure with input layer, output layer and three intermediate Layers

4.2 Genetic Algorithm

GA (Genetic Algorithm) represents the heuristic process and an optimization strategy observed by the procedure of the natural method of selection. It is broadly utilized for finding an optimal solution for the problems related to optimization with huge boundary space. The process of progression of species is mapped, by relying upon naturally inspired things, for example, the crossover process. Moreover, as it doesn't consider the derivatives, it tends to be utilized for both continuous and discrete techniques of optimization[16].

For implementing a Genetic Algorithm, two preconditions must be satisfied, 1) an available representation or characterizing a chromosome and 2) a best suited fitness function to assess created solution representation. In this case, an array of binary value is a genetic type
representation of the solution (Figure 3) and models the (RMSE) Root-Mean-Square Error on the evaluation set which acts as the fitness value. Furthermore, three fundamental tasks that
establish a Genetic Algorithm are the selection, next the crossover and to perform the mutation process applied with heuristic methods.

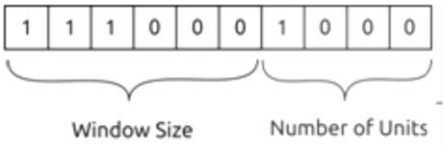

Figure 3: a solution structure as per genetic representation

Figure 4 outlines a flow of complete genetic algorithm procedure, where, beginning solutions(the populations) are generated randomly. Next, the solutions were assessed by a selected fitness function, then the crossover and afterwards the mutation function. This procedure is iterated for a fixed number of times (called the generations in the genetic terminology). Toward the end, an optimal solution with most fitness score could be chosen to consider for best available parameter.

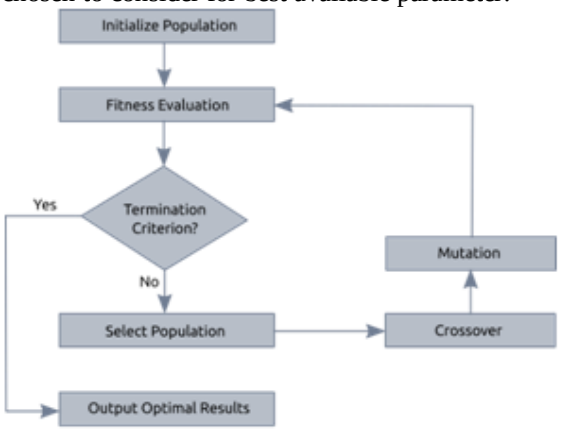

Figure 4: Genetic Algorithm (GA) process flow

4.3 Structure of Proposed Ensemble Model for Covid-19 Forecasting

The vanilla LSTM type of recurrent network system effectively used to address the sequential challenges of data space [17]. It is designed with cells to store the data in the form of blocks which can be connected recurrently.

These defined cells take care of the problem related to vanishing gradient in RNN. Each of the LSTM section contains self-associated cells among the forget gate, output gate and the input gate. These designated gates were intended to store the data state longer than the neural network systems of type feed-forward towards improving performance of the system. A section block in aLSTM model contains cells which is associated recurrently as appeared in Figure 5.

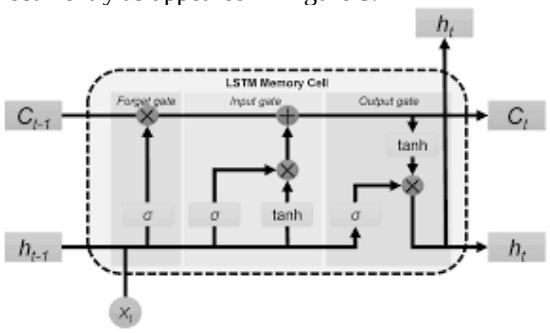

Figure 5: LSTM block outlined by forget gate, input gate, block input, output gate, tangent activation functions

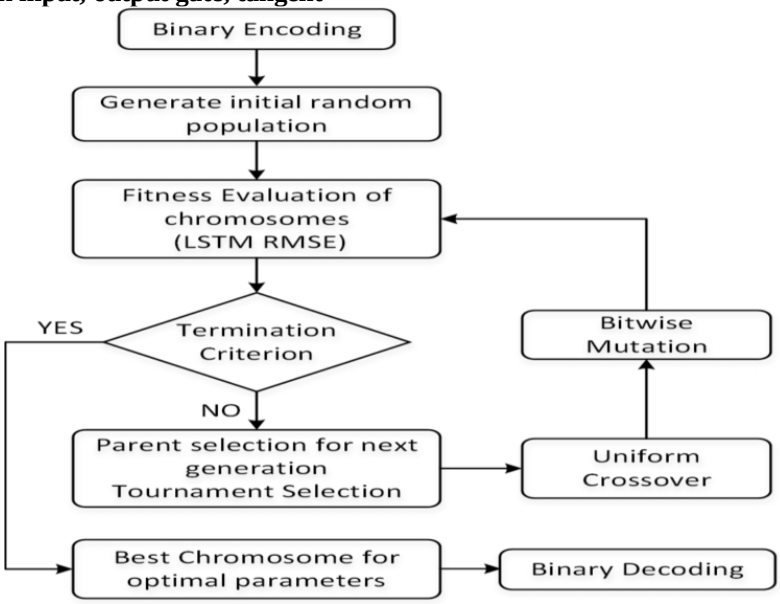

Figure 6: The flow of the Proposed GA-LSTM Methodology for Optimization

The formulas specify the forward pass of a vanilla LSTM implementation:

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

$$i_t = \sigma(W_i \cdot [h_{t-1},\ x_t] + b_i) \quad (2)$$

$$o_t = \sigma(W_o \cdot [h_{t-1},\ x_t] + b_o) \quad (3$$

$$c_t = c_{t-1} \odot f_t + i_t \odot \tanh(W_c \cdot [h_{t-1},\ x_t] + b_c) \quad (4$$

$$h_t = \tanh(c_t) \odot o_t \quad (5$$

GA (Genetic Algorithm) is utilized for finding an optimal window size and no. of units in LSTM based RNN. For the Genetic Algorithm, DEAP python package will be adopted. The basic idea of this methodology is to utilize the algorithm, to discover optimal parameters with the help of two helper methods. The first technique is to fragment the information to make X, Y pair for model preparing. The subsequent technique is to perform three things, a) decoding the algorithm solution to obtain the window size and the number of units. b) Prepare the available data setutilizing window size found by the genetic algorithm and partition into train data and another as the validation data, also c) train the LSTM model with the optimal parameters, compute the RMSE (mean square error) with respect to the validation data and fitness score to be returned as the current solution space to the genetic algorithm. Bernoulli distribution is used for random initialization, similarly, crossover in an ordered form, shuffling the mutation process using theroulette wheel selection technique. Subsequently, the optimal parameters are taken to the model training from the total train set also to test it with the projected test data. The LSTM model is now utilized for covid-19 detection. This neural network has only one hidden layer. To start with, the combined dataset is separated into train and test data with 15% for model learning and 85% for the accuracy test, and the datasets are mixed up. Further the train data to be partitioned as two data groups as 80% in model training and 20% in validation of trained model. Using the keras for implementation the input is taken to covert each element to a vector with the embed layer. The LSTM hidden layer of size 100 is used and utilizing softmax in the final layer to classify the data into output labels. The adam optimzer is adopted and for the estimation of the loss "categorical cross" entropy is utilized.

5. RESULT AND DISCUSSION

5.1 Description of the dataset

The covid-19 is considered from the famous Kaggle repository [18]. This dataset contains India states/union territory's number of confirmed Indian positive cases, confirmed foreign positive cases, cured cases, deaths, and confirmed. The dataset is considered from the year of 2020 – 2021. Totally this dataset is composed of 8 features namely Date, time, State/Union Territory, Confirmed Indian National, Confirmed Foreign National, Cured, Deaths, Confirmed. After the above-mentioned feature engineering approach, the total number of features we got from the considered dataset is 23 features are created based on the date feature and confirmed cases. The feature engineering approach is carried using previously published research paper [19].

Table 1: Number of features obtained by the feature of importance technique and original dataset after the feature engineering approach

Feature of Importance Techniques	Number of Features obtained
Original dataset with feature engineered approach	30
Random Forest (Feature of Importance)	28
Light GBM (Feature of Importance)	27
XG Boost (Feature of Importance)	27

Table 2 gives the evaluation metrics used to analyze the prediction accuracy of the covid-19 dataset using the regressors like Light GBM, XG boost, and Long Short-Term Memory (LSTM) in terms of error rates.

Table 2: Performance Metrics for the Regression techniques used in this research work

Error Name	Equation
Symmetric Mean Absolute Percentage Error (SMAPE)	$\frac{\sum_{t=1}^{n}\lvert F_t - A_t\rvert}{\sum_{t=1}^{n}(F_t + A_t)}$
Mean Absolute Percentage Error (MAPE)	$\frac{1}{n}\sum_{t=1}^{n}\left\lvert\frac{A_t - F_t}{A_t}\right\rvert$
Root Mean Squared Error (RMSE)	$\sqrt{\frac{\sum_{t=1}^{n}(F_t - A_t)^2}{n}}$

Table 3 gives the SMAPE value obtained (in %) by the LGBM, XG Boost, and LSTM regression technique with the feature engineered datasets. From the table 2, it is clear that the features obtained by FOI methods with LSTM gives minimum SMAPE value than the other regressors.

Table 3: SMAPE value obtained (in %) by the LGBM, XG Boost, and LSTM regression technique with the feature engineered datasets

Feature of Importance	SMAPE (in %) obtained by Regression Techniques			
	Proposed Ensemble Model	**LSTM**	**XG Boost**	**Light GBM**
Original Dataset	32.57	43.52	46.85	50.07
RF(FOI)	21.85	36.74	38.52	39.44
LGBM (FOI)	12.96	28.47	29.56	30.45
XG Boost (FOI)	13.47	29.37	30.89	31.48

Table 4 gives the MAPE value obtained (in %) by the LGBM, XG Boost, and LSTM regression technique with the feature engineered datasets. From the table 4, it is clear that the features obtained by FOI methods with LSTM gives minimum MAPE value than the other regressors.

Table 4: MAPE value obtained (in %) by the LGBM, XG Boost, and LSTM regression technique with the feature engineered datasets

Feature of Importance	MAPE (in %) obtained by Regression Techniques			
	Proposed Ensemble Model	**LSTM**	**XG Boost**	**Light GBM**
Original Dataset	36.78	42.78	47.83	50.34
RF(FOI)	24.17	35.44	36.82	37.72
LGBM (FOI)	11.85	28.59	31.84	32.84
XG Boost (FOI)	12.93	29.72	32.86	33.82

Table 5 gives the RMSE value obtained (in %) by the LGBM, XG Boost, and LSTM regression technique with the feature engineered datasets. From the table 5, it is clear that the features obtained by FOI methods with LSTM gives minimum RMSE value than the other regressors.

Table 5: RMSE value obtained (in %) by the LGBM, XG Boost, and LSTM regression technique with the feature engineered datasets

Feature of Importance	RMSE (in %) obtained by Regression Techniques			
	Proposed Ensemble Model	**LSTM**	**XG Boost**	**Light GBM**

Original Dataset	33.14	41.68	45.52	46.64
RF(FOI)	20.58	31.57	36.74	38.87
LGBM (FOI)	11.23	32.75	38.97	40.82
XGBoost (FOI)	10.37	31.55	37.28	39.25

6. CONCLUSION

COVID 19 has spread swiftly over the world because to a lack of a good vaccine; as a result, early detection of persons infected with this virus is critical in attempting to contain it by quarantining afflicted people and providing medical assistance as needed to prevent the virus's spread. Through this research work, an ensemble model is proposed with Deep Learning techniques and Optimization method. The proposed ensemble model is working with proposed feature engineering approach for improving the efficiency of the covid-19 forecast. The performance of the Proposed Ensemble model is evaluated with existing regression models like LSTM, XG Boost, Light GBM for the original dataset, feature of importance techniques (RF, LGBM, XG Boost) applied datasets using different error evaluation metrics like MAPE, SMAPE, and RMSE. From the results obtained, it is clear that the proposed Ensemble model with proposed feature engineering approach and feature of importance methods reduces the MAPE, SMAPE and RMSE than the existing regression models like LSTM, XGBoost and Light GBM regression models.

REFERENCES

[1] Kraemer, Moritz UG, et al. "The effect of human mobility and control measures on the COVID-19 epidemic in China." *Science* 368.6490 (2020): 493-497.

[2] Zu, Zi Yue, et al. "Coronavirus disease 2019 (COVID-19): a perspective from China." *Radiology* (2020): 200490.

[3] Sumner, Andy, Chris Hoy, and Eduardo Ortiz-Juarez. "Estimates of the Impact of COVID-19 on Global Poverty." *UNU-WIDER, April* (2020): 800-9.

[4] Feroze, Navid. "Forecasting the patterns of COVID-19 and Causal Impacts of Lockdown in Top Ten Affected Countries using Bayesian Structural Time Series Models." *Chaos, Solitons & Fractals* (2020): 110196.

[5] Abu-Rayash, Azzam, and Ibrahim Dincer. "Analysis of mobility trends during the COVID-19 coronavirus pandemic: Exploring the impacts on global aviation and travel in selected cities." *Energy research & social science* (2020): 101693.

[6] Tandon, Hiteshi, et al. "Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future." *arXiv preprint arXiv:2004.07859* (2020).

[7] Chandu, Viswa Chaitanya. "Time series forecasting of COVID-19 confirmed cases with ARIMA model in the South East Asian countries of India and Thailand: a comparative case study." *medRxiv* (2020).

[8] Roy, Santanu, Gouri Sankar Bhunia, and Pravat Kumar Shit. "Spatial prediction of COVID-19 epidemic using ARIMA techniques in India." *Modeling Earth Systems and Environment* (2020): 1-7.

[9] Mele, Marco, and Cosimo Magazzino. "Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence." *Environmental Science and Pollution Research* (2020): 1-9.

[10] Ouldali, Naim, et al. "Emergence of Kawasaki disease related to SARS-CoV-2 infection in an epicentre of the French COVID-19 epidemic: a time-series analysis." *The Lancet Child & Adolescent Health* 4.9 (2020): 662-668.

[11] Salgotra, Rohit, Mostafa Gandomi, and Amir H. Gandomi. "Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming." *Chaos, Solitons & Fractals* (2020): 109945.

[12] Guo, Tao, et al. "Cardiovascular implications of fatal outcomes of patients with coronavirus disease 2019 (COVID-19)." *JAMA cardiology* (2020).

[13] Vokó, Zoltán, and János György Pitter. "The effect of social distance measures on COVID-19 epidemics in Europe: an interrupted time series analysis." *Gero Science* (2020): 1-8.

[14] Reddy, K. Shyam Sunder, YCA Padmanabha Reddy, and Ch Mallikarjuna Rao. "Recurrent neural network based prediction of number of COVID-19 cases in India." *Materials Today: Proceedings* (2020).

[15] Zeroual, Abdelhafid, et al. "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study." *Chaos, Solitons & Fractals* 140 (2020): 110121.

[16] Salgotra, Rohit, Mostafa Gandomi, and Amir H. Gandomi. "Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming." *Chaos, Solitons & Fractals* 138 (2020): 109945.

[17] Barman, Arko. "Time series analysis and forecasting of covid-19 cases using LSTM and ARIMA models." *arXiv preprint arXiv:2006.13852* (2020).

[18 https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-india?select=covid_19_india.csv

[19] M. Lalli, M. Pandijothi. “An Efficient Feature Engineering Approach For The Forecasting Of Covid-19.”*Webology*18(4) (2020): 817-827.