JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

COVID-19 forecasting using LSTM neural network

Deepshikha Bansal
Student, Department of CSE
MERI College of Engineering
222db.angel@gmail.com

Neeraj Kumar
Faculty, Department of CSE
MERI College of Engineering
Neeraj.kumar@meri.edu.in

Gaurav Kumar
Faculty, Department of CSE
MERI College of Engineering
Gaurav.kumar@meri.edu.in

Abstract- The COVID-19 epidemic poses a serious threat to civilization and has caused irreparable harm to society. Therefore, making prediction of COVID-19 is crucial. . This study predicts single day forecasting over Indian dataset using LSTM. It is performed over short-term prediction data. It was observed that predicted observation comes out to be little lower than the actual one. The performance of model is analysed using RMSE performance metrics. LSTM model shows accurate prediction with RMSE value of 5855.98 for training set and 2743.6 for testing set. The proposed model shows better results with non-linear data.

Keywords- Prediction, COVID-19, Long-short term memory(RMSE), Root Mean Square Error(RMSE)

I. INTRODUCTION

Since the pandemic outbreak, COVID-19 has spread quickly over many nations and areas of the world. On January 30, 2020, the World Health Organization designated COVID-19 as a Public Health Emergency of International Concern (PHEIC)[1]. The WHO said in a statement that the variations spread more quickly and increase the risk of re-infection. Since January, 2020, world has came across 3 major outbreak waves that includes over 650 million cases, and over 6.6 million people died. According to journal of virology, over 1300 variants have been discovered so far. Among these variants, alpha, beta, gamma, delta and omicron. Latest one is omicron which was discovered in 2022.

The COVID-19 infection produces various symptoms that impact the respiratory system, including fever, cough, illness, diarrhoea, headache, etc[2]. It has a rather high mortality rate (Table.1) and is a very infectious illness. It spreads between people through coming into contact with infected people that have been exposed to viral contamination.

Numerous researches on COVID-19 have predicted the spread of the virus and modelled how many individuals will become infected by it[3].

TABLE.1. COMPARISION OF OUTBREAK IN HISTORY

Outbreak	First discovered in	Mortality rate
Coronavirus (India)	2019(China)	1.8(May,2021)
Ebola	2014(Guinea)	39.5
SARS	2003(China)	9.5
MERS	2012(Saudi Arabia)	34
H1N1	2009(Mexico)	0.25

II. RELATED WORK

A modified SEIR, LSTM model has been used to forecast the epidemic trends of COVID-19 using population data of China [4]. This model predicts several peaks and sizes of epidemics in the intervals or time series fashion. A LSTM and ANFIS model was used in [5] to analyse and predict new cases of COVID-19 in Bangladesh. It was identified that LSTM performed better than ANFIS. Training of ANFIS requires extra time and hardware to predict new cases.

A novel matrix profile based-LSTM attention and other baseline neural network model was develop to predict three indicators i.e. admitted percentage in hospital, confirmed cases and death cases of USA[6]. This matrix profile based model performed better than others model used in the paper. In [7], several models like LSTM, CNN, ANN and ARIMA have been used to predict the COVID-19 infection and analyse the impact on public health of Hubei province of China. It was found that LSTM performance is best among other 4 models with ARIMA performed worst.

VOC-LSTM, GRU, Bi-LSTM model have been used to predict confirmed cases[1] with RMSE value of 0.0343 and R^2 value of 96.83%. VOC-LSTM outperformed in providing higher accuracy in long-term prediction. SIRVD-DL model performed 51% better over other LSTM model to predict daily forecasting of COVID-19 new cases. A SEIR model was used to track movements of people in Census Block Groups in 10 cities of USA using mobility network data[8]. It was found that disadvantages and lower socio-economic groups becomes the super spreader of Covid-19 outbreak. A LASSO and Logistic regression model[9] was used to develop a risk score known as COVID-GRAM framework to predict critical illness of patients of 575 hospitals in 31 provinces of China. This model calculated the risk score of patients based on predicted variables. To develop and external validate the routine blood test of clinical data from 66 hospitals of USA using XGBoost model[10].

A Monte Carlo Markov Chain method in Bayesian network has been develop to predict time-delay adjusted risk for death for COVID-19 daily cases of confirmed and death cases in the Wuhan and Hubei city of China[11]. An ANFIS, MLP-Imperialist Competitive algorithm for optimisation model[12] was used to predict COVID-19 pandemic daily cases and death cases of Hungary. It was found that MLP-ICA shows highest accuracy over other algorithms and possibility of drop of outbreak and rate of mortality. An Exponential smoothing model based on multiplicative error was used to predict 10-days ahead for confirmed cases of COVID-19 in China[13]. It was estimated that there is high probability of drop of outbreak.

A Fbprophet model was used to predict the COVID-19 epidemic trends globally including India using ArcGIS platform data[14]. Overall fit of the actual and predicted data is found to be good. Deep learning based COX model was used to predict COVID-19 risk model based on clinical symptoms of the patients of 575 hospitals of China[15]. It was found that proposed model is statistically better than other two traditional models. Even in follow up stage C-Index and AUC value has improved.

A K-means-LSTM, XGBoost and SEIR model has been used to predict COVID-19 confirmed cases in Louisiana state of USA[16]. The K-means LSTM shows better accuracy than traditional SEIR model. A Fbprophet model[17] was used to predict the COVID-19 cases in Saudi Arabia. The actual data is slightly higher than predicted one. Model is good at forecasting death cases and poor at forecasting recovered cases. A Cauchy exploration strategy with beetle antennae search[18] and ANFIS model has been used to forecast COVID-19 cases in China and the USA. This hybrid model outperformed over other traditional models.

In [19], a VGG16, ResNet50, and Xception stacked model was used to predict COVID-19 from CT-Scan images of patients in a hospital from Sao Paulo, Brazil. The stacked framework outperformed better than individual DL models. A pre-trained VGG19 model, Logistic Regression, XGBoost classifier[20] was used to develop a bi-level classification model which classifies normal Pneumonia and COVID-19 patients of China using Chest X-rays. The bi-level classification model is flexible in terms of choosing different classifiers. Combination of Logistic regression and XGBoost model shows best results in accuracy, recall as compare to other state-of-the-art models on same dataset. A stage wise approach of pre-processing and prediction of COVID-19 due to heart disease of two major hospitals of USA has been performed used Fitness-Dragonfly, Partial Swarm Optimisation, and Grey Wolf Optimisation algorithm[21]. The F-DA model outperformed over PSO, GWO in hidden neuron optimisation framework.

III. DATA COLLECTION AND STUDY AREA

The data was collected by Johns Hopkins University System Science and Engineering Center(CSSE). This data includes confirmed cases, recovered cases and death cases globally. For this paper, India confirmed cases data has been used with available parameters of province, longitude, latitude and cases according to the date. https://github.com/CSSEGISandData/COVID-19 . The time series data considers confirmed cases of India from 22 January, 2020 till today. The daily increase of cases in India is shown in figure 1. This time series data splits data into training and testing set in ratio of 75:25 with 733 and 245 entries.

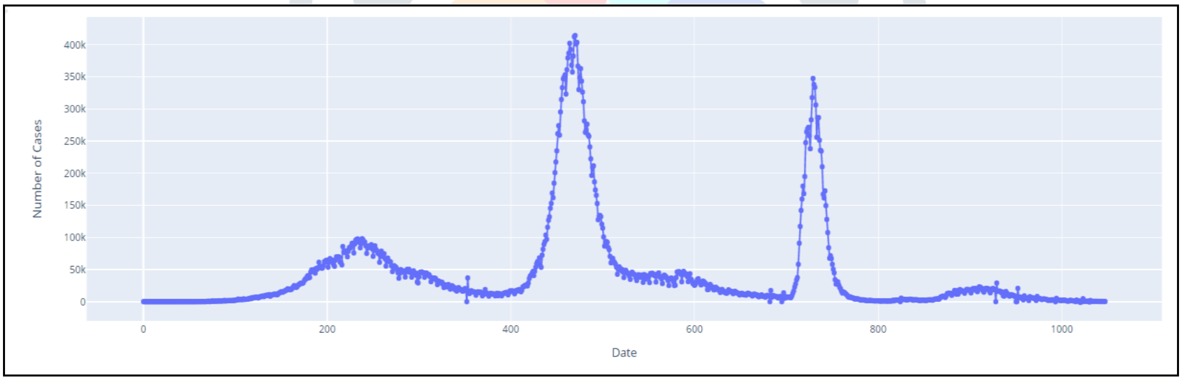
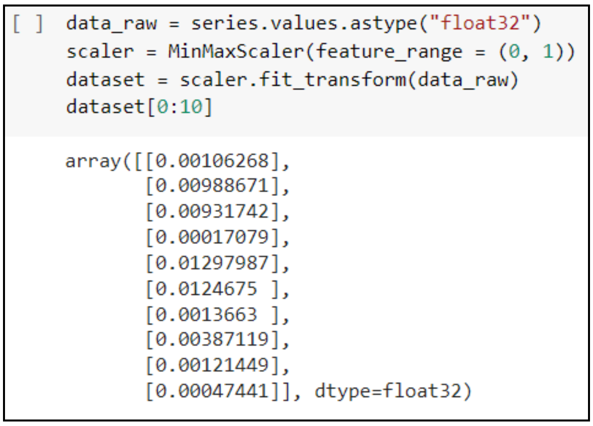


Figure 1. Daily increase of cases in India

Feature scaling is performed over the raw data using Min-Max Scalar in the Scikit-Learn library of python (figure 2), and the data was scaled between 0 to 1.



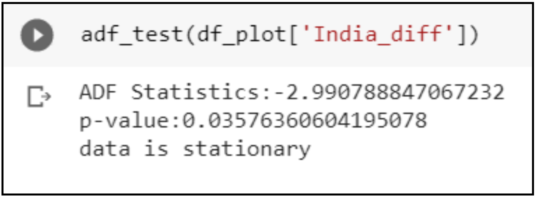
```
[ ]  data_raw = series.values.astype("float32")
     scaler = MinMaxScaler(feature_range = (0, 1))
     dataset = scaler.fit_transform(data_raw)
     dataset[0:10]

     array([[0.00106268],
            [0.00988671],
            [0.00931742],
            [0.00017079],
            [0.01297987],
            [0.0124675 ],
            [0.0013663 ],
            [0.00387119],
            [0.00121449],
            [0.00047441]], dtype=float32)
```

Figure.2 Min-Max Scalar for feature scaling

IV. DESCRIPTIVE STATISTICS

In order to examine stationary or non-stationary, a ADF-Test was applied on the series. Initially time series was non stationary(Figure 3). Shifting of series was performed to make series as stationary for better prediction.

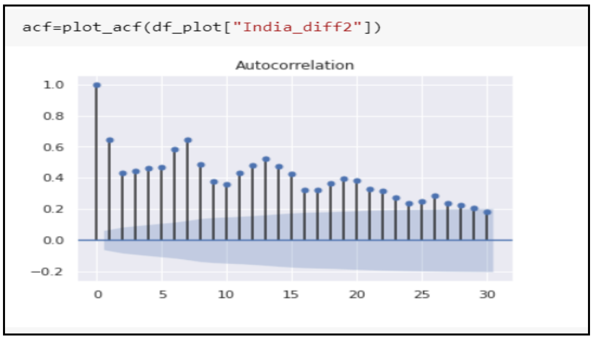


```
adf_test(df_plot['India_diff'])

ADF Statistics:-2.990788847067232
p-value:0.03576360604195078
data is stationary
```

Figure.3 ADF-test for stationary criteria

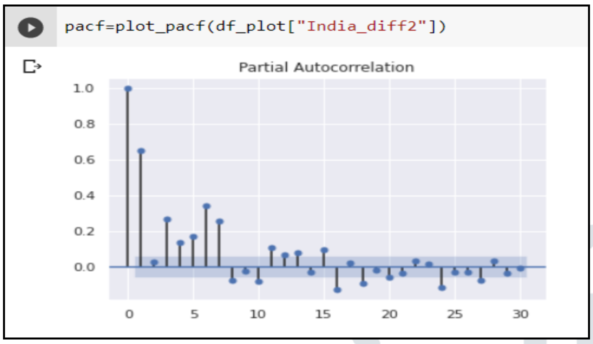
The autocorrelation coefficients of COVID-19 time series vary between -0.2 to +0.2, which indicates that series exhibit non persistence and make data stationary. Further autocorrelation function (ACF) and Partial autocorrelation function are evaluated for the series and are summarized below in Figures 4, 5.



```
acf=plot_acf(df_plot["India_diff2"])
```

Figure 4. Autocorrelation for time series(India)

Statistical summary of the data is depicted in Table 2.



```
pacf=plot_pacf(df_plot["India_diff2"])
```

Figure 5. Partial Autocorrelation for time series(India)

TABLE II. STASTICAL SUMMARY

Statistics	India*(covid-19)*
Count	978
Mean	4031.37
Std	6699.18
Min	0.0
Max	52697.0

V. METHODOLOGY

Over 200 nations were impacted by the global epidemic known as COVID-19. To stop the spread of an outbreak, it is essential to analyse its epidemiological behaviour. The spread of the pandemic in several nations has been predicted using statistical and mathematical modelling techniques. In the study and forecasting of time series data, deep learning (DL) shows promising outcomes. The temporal dependencies and structures in the data, such as trends and seasonality, may be automatically learned by DL models.

ANN is a computational model just like the human brain. It is a simplified model of a biological neuron system[22]. As Neurons can deal with complex, non-linear information of the human brain, in the same way artificial neurons in the neural networks are used to deal with soft computing problems where approximate results can be achieved using approximate models[23]. ANN consists of an input layer, hidden layers, and an output layer. The information is processed through these layers, and the desired result can be achieved at the output layer.

A LSTM model was developed to predict monthly confirmed cases[24] in India with error percentage of less than 10%. A daily and weekly prediction of new cases and deaths using transfer learning LSTM[25] with RMSE of new cases is 0.21 and of deaths is 0.156. In [26], CNN-LSTM model was used to predict daily confirmed cases worldwide as well as for India. RMSE of worldwide was 331.27 and RMSE of India for confirmed cases was 280.19.

In this paper, LSTM is used. LSTM networks are preferably used in non-linear data like time series which uses temporal data to generate meaningful insights on COVID-19 situation. LSTM uses memory units which increase the ability of LSTM to learn dependencies especially long-term. These units permit the network to learn, forget and update hidden layers (figure 6). LSTM intakes various epidemic parameters and provides outputs as predicted figures (numbers).

There is an $X_t = [p_1, p_2, p_3 \ldots\ldots p_9, p_{10}]$ input vector at a given time t and $[p_2, p_3 \ldots\ldots p_9, p_{10}]$ are the weather parameters or predictors. At time t, long-term C_{t-1} and short-term H_{t-1} are updated.

$I_t = \tanh\,(w_{xi}X_t + w_{hi}H_{t-1} + b_i)$, $w_*$are weight matrices, I_t is an input gate, , b_* —biases

$J_t = \mathrm{sigm}\,(w_{xj}X_t + w_{hi}H_{t-1} + b_j)$, J_t is an input moderation gate

$F_t = \mathrm{sigm}\,(w_{xf}X_t + w_{hf}H_{t-1} + b_f)$, $F_t$is an forget gate

$O_t = \tanh\,(w_{xo}X_t + w_{ho}H_{t-1} + b_o)$, O_t is an output gate

$C_t = C_{t-1} \odot F_t + I_t \odot J_t)$

$H_t = \tanh\,(C_t) \odot O_t$, where $\odot$ is an element-wise vector product ,C_t And H_t are two hidden states that allow making decisions for short time period. Switching among $I_t and\ F_t$ gates is used to select current inputs or forgets its earlier memory. Output gate O_t describes memory C_t required to transfer to H_t as hidden state.

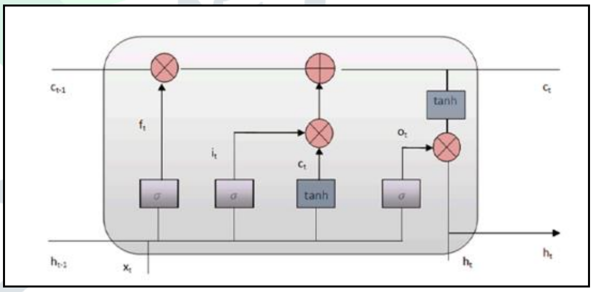


Figure.6 LSTM Architecture[27]

In order to get good results, the LSTM's performance is heavily reliant on the hyper-parameters chosen. For handling the sequences in the input observations, long short-term memory is used. This memory is capable of learning input-to-output function mappings that are not supported by MLP and CNN.

VI. PERFORMANCE ANALYSIS

Comparing the effectiveness of our proposed prediction method using performance measure index, such as the Root Mean Square Error(RMSE) and is elaborated as follows :

$$= \sqrt{\frac{1}{p}\Sigma_{i=1}^{p}(a_i - f_i)^2} \qquad \ldots\ldots(\text{equ.1})$$

p=number of observations; a_i=actual terms; f_i=predicted terms.

It is the error difference of original observations and the forecasted observations. For better results, RMSE value should be smaller. Performance of LSTM model of prediction of COVID-19 confirmed cases is demonstrated using RMSE value.

VII. RESULTS AND DISCUSSION

The time series data considers confirmed cases of India from 22 January, 2020 till today. This time series data splits data into training and testing set in ratio of 75:25 with 733 and 245 entries. In this work, single day prediction is performed over Indian dataset. The series generated after the pre-processing act as an input for the LSTM input layer. A dense layer has used to create neurons in hidden layers. It associates the current layer with the previous layer. This model used three hidden layers and 500, 500, and 400 neurons were created for each of the hidden layer using the dense() function. To reduce over-fitting, the dropout function was used. 0.4 Value is taken as input in the dropout function, i.e., 40% of the neuron will be deactivated while training. Over 300 epochs were used to observe the loss rate(Table 3). The loss rate depicted is below the 0.0153 value.

TABLE.3 PARAMETERS AND POSSIBLE VALUES

Epochs	300
Batch Size	32
Optimizer	Adam
Activation function	tanh
Loss	Mean Square Error

RMSE of training dataset comes out to be 5855.98 whereas RMSE of testing dataset comes out to be 2743.86(Table 4).

TABLE.4 COMPARING TRAINING & TESTING SET

Summary	**Training**	**Testing**
Steps	23	8
Time/Steps	29 m/s	27 m/s
RMSE	5855.98	2743.86

A graph is also plotted to compare the actual value and the predicted value for both the training set and test set(figure 7).

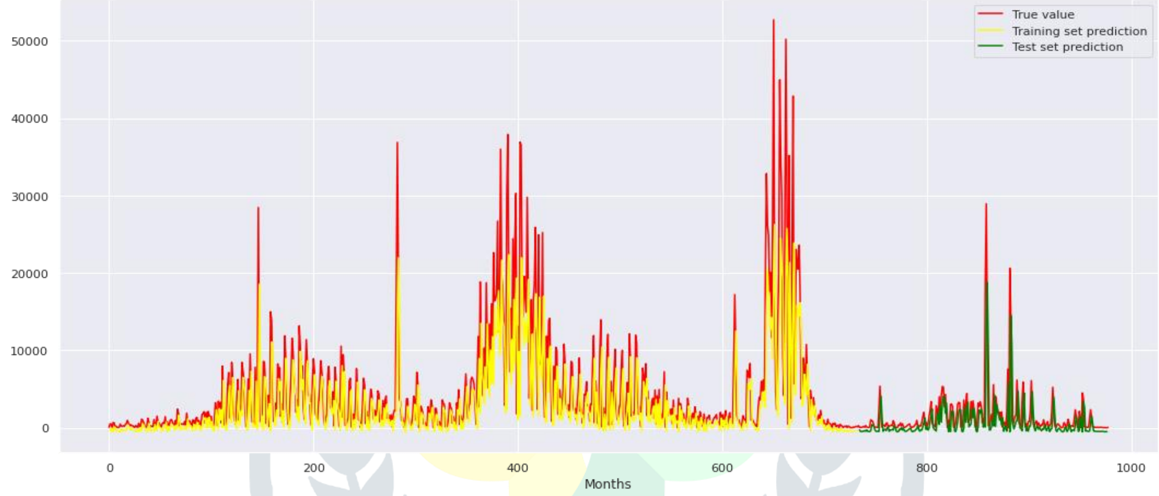


Figure.7 Graph of COVID-19 over predicted training and testing set Vs Actual set.

VIII. CONCLUSIONS

The COVID-19 epidemic poses a serious threat to civilization and has caused irreparable harm to society. By forecasting pandemic breakouts and developing vaccinations, research is being done to reduce the number of fatalities among humans. In the present environment, accurate COVID-19 prediction using deep learning has drawn increased interest. Deep learning techniques are increasingly important for properly tackling non-linear problems. This study predicts single day forecasting over Indian dataset using LSTM. It is performed over short-term prediction data. It was observed that predicted observation comes out to be little lower than the actual one. LSTM model shows accurate prediction (figure.7) with RMSE value of 5855.98 for training set and 2743.6 for testing set. For the future work,

1. Recommendation of hybrid deep learning model can be used for prediction.
2. For more accuracy, need to apply dataset on different deep learning models.
3. Incorporate optimisation mechanism to increase capability of the model.
4. Wise selection of parameters of the given data and increase the parameters count for accuracy.
5. Extension of the proposed model on long-term prediction.

IX. REFRENCES

[1] Z. Liao, Y. Song, S. Ren, X. Song, X. Fan, and Z. Liao, "VOC-DL : Deep Learning Prediction Model for COVID-19 Based on VOC Virus Variants," *Comput. Methods Programs Biomed.*, vol. 224, p. 106981, 2022, doi: 10.1016/j.cmpb.2022.106981.

[2] A. H. Elsheikh *et al.*, "Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia," *Process Saf. Environ. Prot.*, vol. 149, pp. 223–233, 2021, doi: 10.1016/j.psep.2020.10.048.

[3] Z. Liao, P. Lan, X. Fan, B. Kelly, A. Innes, and Z. Liao, "SIRVD-DL : A COVID-19 deep learning prediction model based on time-dependent SIRVD," *Comput. Biol. Med.*, vol. 138, no. August, p. 104868, 2021, doi: 10.1016/j.compbiomed.2021.104868.

[4] Z. Yang *et al.*, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *J. Thorac. Dis.*, vol. 12, no. 3, pp. 165–174, 2020, doi: 10.21037/jtd.2020.02.64.

[5] A. A. Chowdhury, K. T. Hasan, and K. K. S. Hoque, "Analysis and Prediction of COVID-19 Pandemic in Bangladesh by Using ANFIS and LSTM Network," *Cognit. Comput.*, vol. 13, no. 3, pp. 761–770, 2021, doi: 10.1007/s12559-021-09859-0.

[6] Q. Liu, D. L. X. Fung, L. Lac, and P. Hu, "A Novel Matrix

Profile-Guided Attention LSTM Model for Forecasting COVID-19 Cases in USA," *Front. Public Heal.*, vol. 9, no. October, 2021, doi: 10.3389/fpubh.2021.741030.

[7] W. Wang, J. Cai, J. Xu, Y. Wang, and Y. Zou, "Prediction of the COVID-19 infectivity and the sustainable impact on public health under deep learning algorithm," *Soft Comput.*, vol. 0, 2021, doi: 10.1007/s00500-021-06142-0.

[8] S. Chang *et al.*, "Mobility network models of COVID-19 explain inequities and inform reopening," *Nature*, vol. 589, no. 7840, pp. 82–87, 2021, doi: 10.1038/s41586-020-2923-3.

[9] W. Liang *et al.*, "Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19," *JAMA Internal Medicine*, vol. 180, no. 8. pp. 1081–1089, 2020, doi: 10.1001/jamainternmed.2020.2033.

[10] T. B. Plante *et al.*, "Development and external validation of a machine learning tool to rule out COVID-19 among adults in the emergency department using routine blood tests: A large, multicenter, real-world study," *J. Med. Internet Res.*, vol. 22, no. 12, pp. 1–12, 2020, doi: 10.2196/24048.

[11] E. Shim, K. Mizumoto, W. Choi, and G. Chowell, "Estimating the risk of COVID-19 death during the course of the outbreak in Korea, february–may 2020," *J. Clin. Med.*, vol. 9, no. 6, pp. 1–9, 2020, doi: 10.3390/jcm9061641.

[12] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen, "COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3590821.

[13] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID-19," *PLoS One*, vol. 15, no. 3, pp. 1–8, 2020, doi: 10.1371/journal.pone.0231236.

[14] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," *Chaos, Solitons and Fractals*, vol. 139, p. 110058, 2020, doi: 10.1016/j.chaos.2020.110058.

[15] W. Liang *et al.*, "Early triage of critically ill COVID-19 patients using deep learning," *Nat. Commun.*, vol. 11, no. 1, pp. 1–7, 2020, doi: 10.1038/s41467-020-17280-8.

[16] S. R. Vadyala, S. N. Betgeri, E. A. Sherer, and A. Amritphale, "Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM," *Array*, vol. 11, p. 100085, 2021, doi: 10.1016/j.array.2021.100085.

[17] M. Khayyat, K. Laabidi, N. Almalki, and M. Al-zahrani, "Time Series Facebook Prophet Model and Python for COVID-19 Outbreak Prediction," pp. 0–12, 2021, doi: 10.32604/cmc.2021.014918.

[18] M. Zivkovic, N. Bacanin, K. Venkatachalam, and A. Nayyar, "COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach," vol. 66, no. November 2020, 2021.

[19] S. Biswas, S. Chatterjee, A. Majee, S. Sen, and F. Schwenker, "applied sciences Prediction of COVID-19 from Chest CT Images Using an Ensemble of Deep Learning Models," 2021.

[20] S. Das, S. Deep, S. Malakar, J. D. Velásquez, and R. Sarkar, "Bi-Level Prediction Model for Screening COVID-19 Patients Using Chest X-Ray Images," *Big Data Res.*, vol. 25, p. 100233, 2021, doi: 10.1016/j.bdr.2021.100233.

[21] S. Koppu, P. Kumar, R. Maddikunta, and G. Srivastava, "Deep learning disease prediction model for use with intelligent robots," *Comput. Electr. Eng.*, vol. 87, p. 106765, 2020, doi: 10.1016/j.compeleceng.2020.106765.

[22] S. Haykin and N. Network, "A comprehensive foundation," *Neural Networks*, vol. 2, no. 2004, p. 41, 2004.

[23] D. R. Nayak, A. Mahapatra, and P. Mishra, "A Survey on Rainfall Prediction using Artificial Neural Network," *Int. J. Comput. Appl*, vol. 72, no. 16, pp. 32–40, 2013.

[24] A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," *Sci. Total Environ.*, vol. 728, p. 138762, 2020, doi: 10.1016/j.scitotenv.2020.138762.

[25] Y. Gautam, "Transfer Learning for COVID-19 cases and deaths forecast using LSTM network," *ISA Trans.*, vol. 124, pp. 41–56, 2022, doi: 10.1016/j.isatra.2020.12.057.

[26] S. Ketu, P. Kumar, and P. K. Mishra, "India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability," *Soft Comput.*, vol. 26, no. 2, pp. 645–664, 2022, doi: 10.1007/s00500-021-06490-x.

[27] J. Devaraj *et al.*, "Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?," *Results Phys.*, vol. 21, no. January, p. 103817, 2021, doi: 10.1016/j.rinp.2021.103817.