IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.8, Issue.6, June 2020 E-ISSN: 2347-2693

Forecasting novel COVID-19 confirmed cases in India using Machine Learning Methods

Saroj S. Date[1*], Sachin N. Deshmukh[2]

[1]Dept. of Computer Science & Engineering, Jawaharlal Nehru Engineering College, MGM University, Aurangabad (MS), India

[2]Dept. of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS) India

**Corresponding Author: sarojdate@jnec.ac.in*

DOI: https://doi.org/10.26438/ijcse/v8i6.5762 | Available online at: www.ijcseonline.org

Received: 26/May/2020, Accepted: 20/June/2020, Published: 30/June/2020

Abstract**—** Nowadays, there is a very adverse impact on economic, cultural, social and almost all fields in the world because of Covid-19. The Covid-19 term is described as -'CO' for corona, 'VI' for virus, and 'D' for disease. It is an infectious disease caused by severe acute respiratory syndrome which is transmitted through respiratory droplets and contact routes. Since December 2019, corona-virus disease (COVID-19) has out-broke from the country China. Till now, more than 78, 23, 289 people are infected and more than 4 Lakhs of deaths have been caused worldwide. Unfortunately, the number of infections and deaths are still increasing rapidly which has put the world in a different state. Artificial Intelligence can play a key role to infection forecasting in national and provincial levels in many countries. The objective of this study is to use machine learning methods to forecast the number of cases for the next 2 weeks, i.e. till 30th June 2020. The data was collected from 22nd January to 15th June 2020 by nationally recognized sources. The data file contains the cumulative count of confirmed, death and recovered cases of COVID-19 from different countries from the date 22nd January 2020.In this study, the outbreak of this disease has been analyzed for India till 15th June 2020 and predictions have been made for the number of cases for the next two weeks**.**

Keywords—Covid-19, Corona, Corona Virus, Machine Learning, Forecasting, Artificial Intelligence, time series forecasting

I. INTRODUCTION

On 31st December 2019, the novel Corona virus, known as COVID-19 was reported in Wuhan, China for the very first time. Corona viruses are the infectious virus which has adverse affect on the respiratory system of humans. The symptoms of COVID-19 may or may not be visual in infected individual, therefore the spread rate can be faster. Till now, effective and well-tested vaccine against CoVID-19 has not been invented, only precautions are the safety measures. Though the continuous efforts are going on , the virus has managed to spread in most of the territories in the world and World Health Organization (WHO) has announced COVID-19 as Pandemic. Most of the countries in the world are working cooperatively and openly to bring this situation under control.

Data scientists and data mining researchers can play an important role during these types of situations. They can integrate the related data and technology to better understand the virus and its characteristics, which can help in taking right decisions and concrete plan of actions.

As per the daily situation report of WHO, as on 15th June 2020 the COVID-19 transmission scenario reports 78,23,289 confirmed cases with 4,31,541 deaths globally.

Data mining is a technology, developing with database as well as artificial intelligence. It is a processing procedure of extracting credible and effective novel techniques and understandable patterns from the database [1]. Artificial intelligence (AI) is a field of programming building which gives PCs an ability to learn without being unequivocally modified [2]. AI models can be used for estimating and predicting spread rate, so AI is one of the beneficial tools to fight against pandemic like COVID-19.

The forecasting analysis is done by using the algorithms like ANN and time-series [3]. In this paper we are using, time-series algorithm. According to K. Krishna Rani Samal et al., approaches like SARIMA and Prophet can be used for forecasting based on historical data. They concluded that both the SARIMA and prophet model provides a good quality of accuracy. However, the best approach is the prophet model on log transformation which has the least minimum RMSE, MSE value [4]. This model is developed by Facebook, available in python and R. The main contribution of this research paper is forecasting of COVID-19 for the next two weeks i.e. till 30th June 2020 using Prophet Model. In this study, the data was analyzed from 22nd January 2020 to 15th June 2020.

Rest of the paper is organized as follows, Section II contain the related work of various studies carried out for Covid-19 prediction , Section III contain the methodology with flow chart used to carry out the proposed work, Section IV contain the results and discussions after implementation of the proposed model. Finally, section V concludes the paper with future work.

II. RELATED WORK

At the present time, lots of researchers are working on this and similar type of areas.

Upendra Kumar Tiwari & Rizwan Khan have tried to use the machine learning to analyse the current situation created by covid-19 and what may be its impact in future days. They have analyzed that the case of covid-19 in India is going to be same as in Italy or South Korea. India might be going to face its worst days in future if we look the pattern of these countries and India [5].

Herlawati tries to use a soft computing algorithm to predict the pattern of the COVID-19 pandemic in Indonesia. Support Vector Regression was used in Google Interactive Notebook with some kernels for comparison, i.e. radial basis function, linear and polynomial [6].

Dutta,Shawni , Samir Kumar Bandyopadhyay ,Tai-Hoon kim attempted to use Machine Learning Approach to build up model which will help clinical doctors for verification of disease within short period of time and also the paper attempts to predict growth of the disease in near future in the world. Experimental results indicate that the combined CNN-LSTM approach outperforms well over the other model [7].

Rustam et al. demonstrated the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. Four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study prove it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic [8].

Ranjan, Rajesh used susceptible-infected-recovered (SIR) models based on available data to make short and long-term predictions on a daily basis. Based on the SIR model, it is estimated that India will enter equilibrium by the end of May 2020[9].

Fong et al. demonstrated an optimized forecasting model that is constructed from a new algorithm, namely polynomial neural network with corrective feedback (PNN+cf) is able to make a forecast that has relatively the lowest prediction error. The results showcase that the newly proposed methodology and PNN+cf are useful in generating acceptable forecast upon the critical time of disease outbreak when the samples are far from abundant [10].

Petropoulos, Fotios & Makridakis, Spyros introduced an objective approach to predicting the continuation of the COVID-19 using a simple, but powerful method to do so. Assuming that the data used is reliable and that the future will continue to follow the past pattern of the disease, their forecasts suggest a continuing increase in the confirmed COVID-19 cases with sizable associated uncertainty. The risks are far from symmetric as underestimating its spread like a pandemic and not doing enough to contain it is much more severe than overspending and being over careful when it will not be needed. This paper also describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides objective forecasts for the confirmed cases of COVID-19 [11].

Zheng N, Du S, Wang J, et al. proposed a hybrid artificial-intelligence (AI) model for COVID-19 prediction. The experimental results on the epidemic data of several typical provinces and cities in China showed that individuals with corona virus have a higher infection rate within the third to eighth days after they were infected, which is more in line with the actual transmission laws of the epidemic [12].

Heni Bouhamed used a Deep Learning nested sequence prediction models with Long Short-Term Memory (LSTM) architecture for the continuous monitoring of the infection and recovering processes. This model was built based on the epidemic data evolution of 79 countries between the date of their first case and March 13, 2020. The data is based on 12 variables for cumulative case number prediction and 13 variables (among which the cumulative number of cases) for cumulative recoveries number prediction [13].

III. METHODOLOGY

The proposed methodology consists of six steps, as shown in figure 1:
i) Data Collection
ii) Data Cleaning
iii) Data Visualization
iv) Building the model
v) Training the model and
vi) Forecasting

Data Collection
Time series data collected from Kaggle has been used for the experimental result analysis. The time period of data is from 22/01/2020 to 15/06/2020. The data includes the cumulative count of confirmed, death and recovered cases

of COVID-19 from different countries. However, this paper focuses only on India's data for analysis and forecasting of COVID-19 confirmed cases.

Data Preprocessing

This collected data undergoes various steps of pre-processing which makes it more sensible. Data are pre-processed by eliminating missing values, irrelevant values.

Prediction Algorithm

In machine learning, various time series forecasting models are available like ARIMA, SARIMA, GARCH, Dynamic linear models, TBATS, Prophet, LSTM, etc.

Here we are using Prophet. Prophet is forecasting model which allows dealing with multiple seasonalities. It is open source software and is released by Facebook's Core Data Science team. The prophet model assumes that time series can be decomposed as follows:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

The three terms $g(t)$, $s(t)$ and $h(t)$ correspond respectively to trend, seasonality and holiday. The last term is the error term.

Data Analysis and Outbreak Prediction

For analysis and forecasting of number of COVID-19 patients in India, the framework shown in figure 1 is used. The data is collected for the duration of 22nd January 2020 till 15th June 2020 from Kaggle. This dataset has everyday level data on the number of influenced cases, recovery, and deceased. It has a total of 38107 rows and 10 columns. These are the columns in the dataset: Province/State, Country/Region, Lat, Long, Date, Confirmed, Deaths, Recovered, Active and WHO Region.

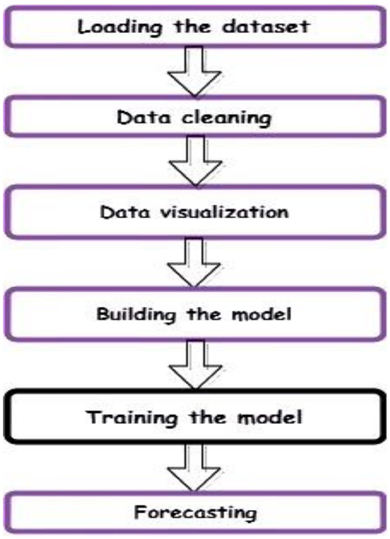


Figure 1. Workflow of the Analysis and Forecasting

We analyzed and visualized the spreading of the virus country-wise as well as globally during the last five months with confirmed cases, recovered cases and deceased. Finally, we predicted the expansion of the virus globally with the help of plotly and prophet python library. In section IV, we have shown the results after visualization.

IV. RESULTS AND DISCUSSION

The aim of this work is to predict the effect of the covid-19 in coming days. To get the results, we have planned the work in following parts:

i) Study the ongoing condition in India
2) Make the analysis from some graphs.
3) Predicting the COVID-19 cases in India using Prophet.

With the help of the collected data some visualization are carried out.

Figure 2 shows a plot. It is about all the countries with active cases of corona-virus disease as on 15th June 2020. The countries having more number of active cases are shown with dark shades.

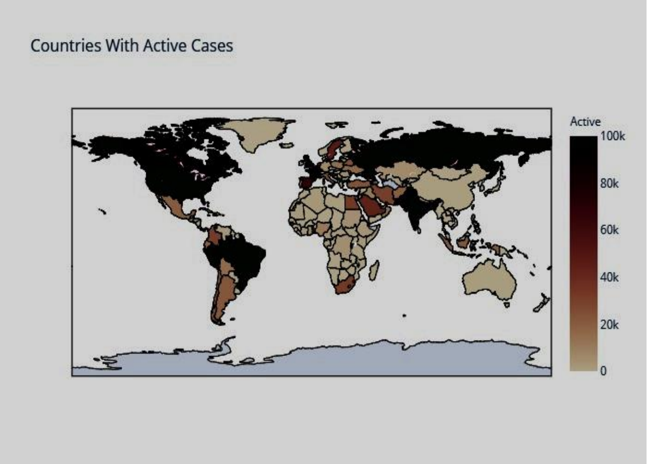
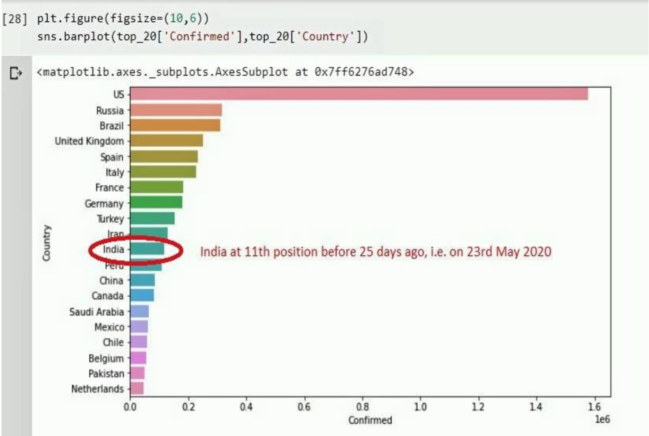


Figure 2. World map showing the countries with active cases as on 15/06/2020

After that, topmost 20 countries with covid-19 confirmed cases are visualized. The study shows that before a month ago, India was at 11th position and on 15th June India is at 4th position.



```
[28] plt.figure(figsize=(10,6))
     sns.barplot(top_20['Confirmed'],top_20['Country'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff6276ad748>
```

India at 11th position before 25 days ago, i.e. on 23rd May 2020

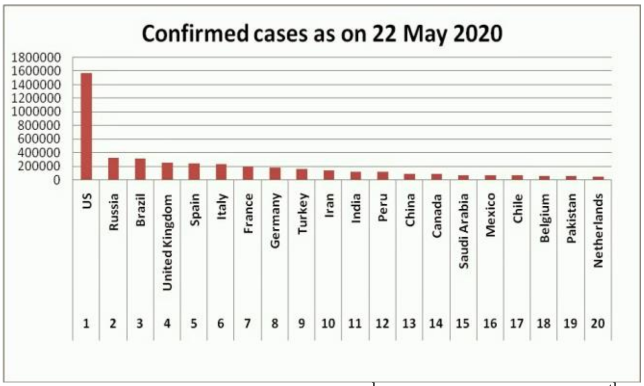
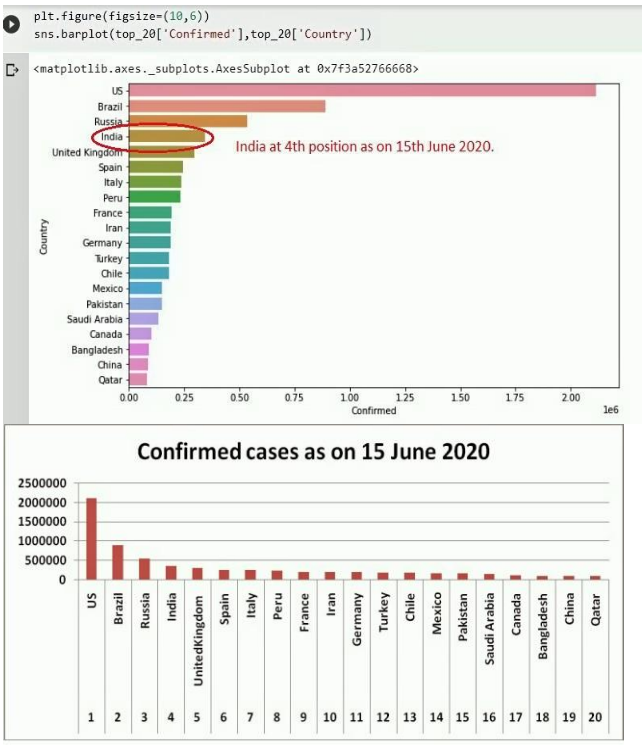


Figure 3. Confirmed cases as on 22nd May, 2020. (India at 11th Position)

Following figure shows the graph containing the number of confirmed cases as on 15th June 2020.



```
plt.figure(figsize=(10,6))
sns.barplot(top_20['Confirmed'],top_20['Country'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3a52766668>
```

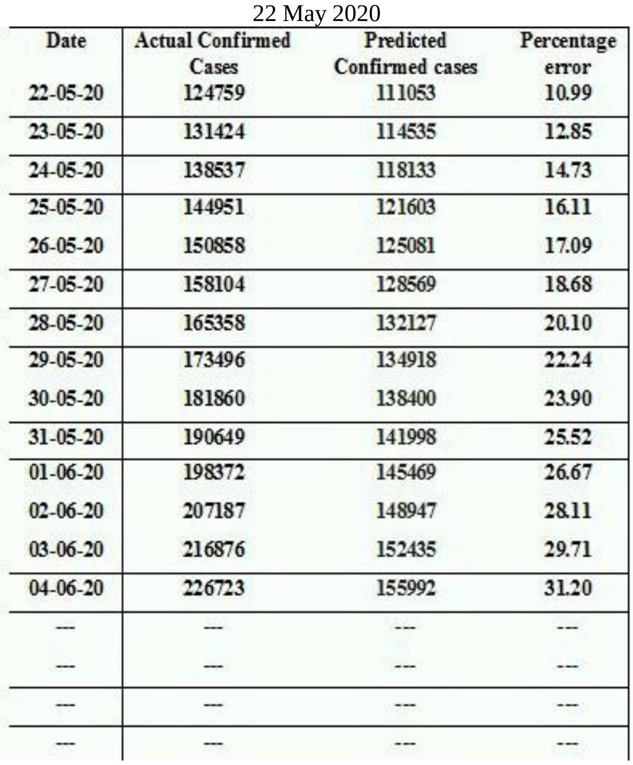
Figure 4. Confirmed cases as on 15th June , 2020. (India at 4th Position*)*

From the above two graphs, we come to know that in India, the number of confirmed cases are increasing day by day because of Covid-19 disease. So, the objective of this work is to create a model that will predict the effect of the corona virus in the next 2 weeks i.e. till 30th June 2020 by using Prophet.

Prophet is the time series forecasting algorithm for future prediction and an adaptive model which uses nonlinear data to predict for yearly, monthly, and daily excluding holiday.

A study was carried out a month ago, i.e.in May 2020, using prophet time series forecasting algorithm, to predict the confirmed cases. Today, when we compare the actual confirmed cases with the predicted confirmed cases, it shows the more than 11% percentage error, as shown in the Table 1.

Table 1 Percentage error calculation for the prediction made on 22 May 2020



Date	Actual Confirmed Cases	Predicted Confirmed cases	Percentage error
22-05-20	124759	111053	10.99
23-05-20	131424	114535	12.85
24-05-20	138537	118133	14.73
25-05-20	144951	121603	16.11
26-05-20	150858	125081	17.09
27-05-20	158104	128569	18.68
28-05-20	165358	132127	20.10
29-05-20	173496	134918	22.24
30-05-20	181860	138400	23.90
31-05-20	190649	141998	25.52
01-06-20	198372	145469	26.67
02-06-20	207187	148947	28.11
03-06-20	216876	152435	29.71
04-06-20	226723	155992	31.20
---	---	---	---
---	---	---	---
---	---	---	---
---	---	---	---

Following figure 5 shows the graphical representation of actual and predicated data.

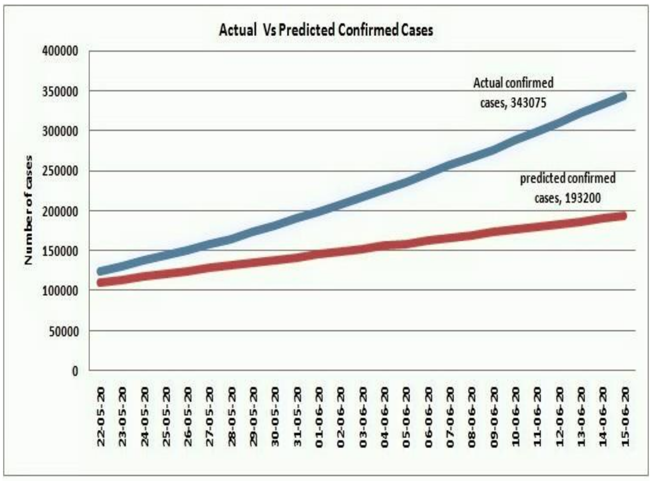
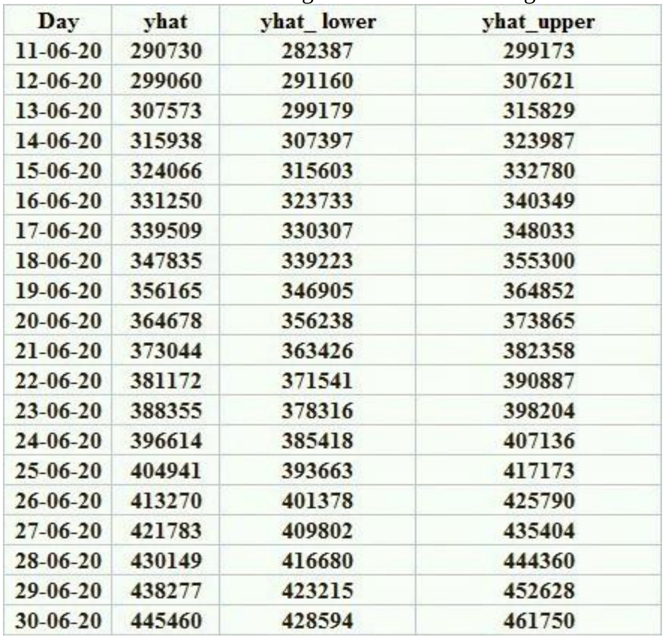


Figure 5: Actual date (15/6/2020) Vs predicted date (22/5/2020) confirmed cases

From the above figure it is clear that, the numbers of actual cases are increasing exponentially.

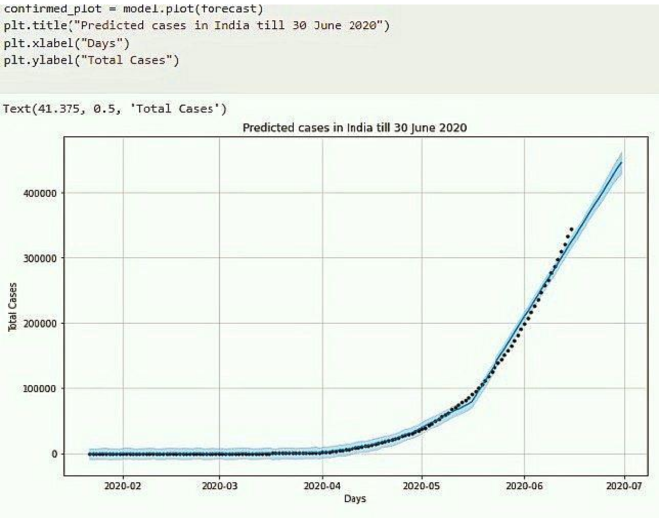
At the end of this paper, following table shows the predictions of number of Confirmed cases in India for the next two week.

Table 2.Prediction of number of COVID-19 patients in India for next two weeks using time series forecasting



Day	yhat	yhat_ lower	yhat_upper
11-06-20	290730	282387	299173
12-06-20	299060	291160	307621
13-06-20	307573	299179	315829
14-06-20	315938	307397	323987
15-06-20	324066	315603	332780
16-06-20	331250	323733	340349
17-06-20	339509	330307	348033
18-06-20	347835	339223	355300
19-06-20	356165	346905	364852
20-06-20	364678	356238	373865
21-06-20	373044	363426	382358
22-06-20	381172	371541	390887
23-06-20	388355	378316	398204
24-06-20	396614	385418	407136
25-06-20	404941	393663	417173
26-06-20	413270	401378	425790
27-06-20	421783	409802	435404
28-06-20	430149	416680	444360
29-06-20	438277	423215	452628
30-06-20	445460	428594	461750

As shown in the table, the number of confirmed cases in India will be 4,45,460 by the end of June 2020.
Following figure 6 shows the predicted cases in India till 30th June 2020.



```
confirmed_plot = model.plot(forecast)
plt.title("Predicted cases in India till 30 June 2020")
plt.xlabel("Days")
plt.ylabel("Total Cases")
```

```
Text(41.375, 0.5, 'Total Cases')
```

Figure 6 Prediction using time series algorithm

The data is analyzed till 15th June 2020 and Covid-19 cases will increase rapidly in next two weeks. Our country might be going to face the worst situation, if we look at the numbers in above table.

V. CONCLUSION AND FUTURE SCOPE

From above findings, we conclude that numbers of Covid-19 Confirmed cases are increasing rapidly. The idea behind this work is to make prediction about the numbers of cases in the near future. Prophet predictive analytics algorithm on Kaggle dataset is used for making predictions. The predictions shows that the confirmed COVID-19 infected cases would be approximately 4,45,460 at the end of June 2020. We hope that these predictions will be helpful in different sectors to take necessary actions. This study will be enhanced in the future. We plan to explore and use the most accurate and appropriate Machine Learning methodologies for forecasting real-time data. Also we can take data at regular intervals automatically with the help of some tool and predict number of cases weekly or at some regular intervals.

REFERENCES

[1] R.S. Walse, G.D. Kurundkar, P. U. Bhalchandra, "*A Review: Design and Development of Novel Techniques for Clustering and Classification of Data*," International Journal of Scientific Research in Computer Science and Engineering, Vol.**6**, Issue.**1**, pp.19-22, **2018**

[2] Hemant Kumar Soni, "*Machine Learning â€“ A New Paradigm of AI*," International Journal of Scientific Research in Network Security and Communication, Vol.**7**, Issue.**3**, pp.31-32, **2019**

[3] Amogha A.K., "*Load Forecasting Algorithms with Simulation & Coding*," International Journal of Scientific Research in Network Security and Communication, Vol.**7**, Issue.**2**, pp.15-20, **2019**

[4] K. Krishna Rani Samal, Korra Sathya Babu, Santosh Kumar Das, Abhirup Acharaya , *"Time Series based Air Pollution Forecasting using SARIMA and Prophet Model*" , In the Proceedings of ITCC 2019: International Conference on Information Technology and Computer Communications, pp **80-85, 2019.**

[5] Upendra Kumar Tiwari & Rizwan Khan, " *Role of Machine Learning to Predict the Outbreak of Covid-19 in India*", Journal of Xi'an University of Architecture & Technology, Vol.**12**, Issue.**4**, pp. 2663-2669, **2020**.

[6] Herlawati ,*"COVID-19 Spread Pattern Using Support Vector Regression*", PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic Journal , Vol.**8**, Issue.**1**, pp. 67-74, **2020**

[7] Dutta,Shawni , Samir Kumar Bandyopadhyay ,Tai-Hoon kim , "*CNN-LSTM Model for Verifying Predictions of Covid-19 Cases*", Asian Journal of Computer Science and Information Technology, Vol.**5**, Issue.**4**, pp. 25-32, **2020**

[8] Rustam, Furqan & Reshi, Aijaz & Mehmood, Arif & Ullah, Dr. Saleem & On, Byungwon & Aslam, Waqar & Choi, Gyu Sang, "*COVID-19 Future Forecasting Using Supervised Machine Learning Models*", in IEEE Access, Vol. **8**, pp. 101489-101499, **2020**

[9] R. Ranjan, *"Predictions for COVID-19 outbreak in India using Epidemiological models*" medRxiv 10.1101/2020.04.02.20051466, **2020**

[10] Simon James Fong, Gloria Li, Nilanjan Dey, Rubén González Crespo, Enrique Herrera-Viedma, " *Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak*", International Journal of Interactive Multimedia and Artificial Intelligence, Vol.**6**, Issue.**1**, pp. 132-140 , **2020**

[11] Petropoulos F, Makridakis S, "*Forecasting the novel coronavirus COVID-19*" , PLOS ONE journal , March 31, 2020. https://doi.org/10.1371/journal.pone.0231236 , **2020**

[12] Zheng N, Du S, Wang J, Zhang H, Cui W, Kang Z, et al. , "*Predicting COVID-19 in China Using Hybrid AI Model*", IEEE Trans Cybern, https://doi.org/10.1109/TCYB.2020.2990162, 2**020**

[13] Heni Bouhamed, "*Covid-19 Cases and Recovery Previsions with Deep Learning Nested Sequence Prediction Models with Long Short-Term Memory (LSTM) Architecture*," International Journal of Scientific Research in Computer Science and Engineering, Vol.**8**, Issue.**2**, pp.10-15, **2020**

Authors Profile

Ms. Saroj S. Date pursued Bachelor of Engineering from SGGS college of Engg. & Tech. , Swami Ramanand Teerth Marathwada University, Nanded in 2004 and Master of Engineering from Dr.Babasaheb Ambedkar Marathwada Univesity, Aurangabad in year 2012. She is currently working as Assistant Professor in Department of Computer Science and Engineering, Jawaharlal Nehru Engineering College, Mahatma Gandhi Mission University, Aurangabad (MS) since 2007. She has published 5 research papers in reputed international journals and IEEE conference and it's also available online. Her main research work focuses on Natural Language Processing, Data Mining and Text Mining. She has 12 years of teaching experience.

Dr. Sachin N. Deshmukh, is currently working as a Professor in Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. He has more than Twenty Five years of experience in teaching. He is member of various professional societies like CSI,IEEE,IETE , ISCA.

He has handled different responsibilities like Director of University Network & Information Center (UNIC) and Director of Internal Quality Assurance Cell (IQAC) in Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. He was Director (Center for Vocational Education and Training), Chief Coordinator of Spoken Tutorial Project of IITB and PET-2014 Coordinator. He has published more than 85 Research papers in reputed International Journals , Conferences and it's also available online His main research work focuses on Data Mining, Text Mining, Social Media Mining, Sentiment Analysis , Big Data Anlytics. For more information visit: http://www.profsnd.in/