COVID-19 RECOVERY CASES PREDICTION USING MACHINE LEARNING

A. Pandian* Raghu Baskaran, Adhithya Venkatesh

[1]*Department of Computer Science and Engineering, SRM University, Kattankulathur, Tamil Nadu*
[2]*Department of Computer Science and Engineering, SRM University, Kattankulathur, Tamil Nadu*
[3]*Department of Computer Science and Engineering, SRM University, Kattankulathur, Tamil Nadu*
pandiana@srmist.edu.in

Abstract

Coronavirus or COVID-19 is a relatively new pandemic gravely affecting the globe. All around the globe people are getting infected by it. Unfortunately a large number of patients who are elderly or immune-compromised die after contracting the virus. In this paper, we aim to test many machine learning techniques and algorithms in order topredict the number of people successfully recovering after contracting the deadly virus. And that prediction is made on a day to day basis. We implemented Random Forest Regression model for an accurate prediction. This study will accurately predict the number of recovered patients on a day to day basis. Thus, we can see a lot of probable uses for this study in the future sense.

Keywords: COVID-19, Machine Learning, Decision Tree Classifier, Random Forest Regression, Random Forest Classifier, Accurate model, Projection

1. Introduction

In Early 2019, the first case of the novel coronavirus was found in the Wuhan province, China. With early symptoms being cough, cold, fever, fatigue. Currently the total infected stand at 250 million cases. Every nation has been alerted by WHO and follows precautions such as lock-down followed by quarantine for a couple of weeks, With social distancing and self-isolation being the new norm. India reported its first case on 30th January, 2020; subsequently the confirmed cases have gradually shot-up to 34 million confirmed cases as of November, 2021.

The preciseness of prediction and its misdoubt depend on the hypotheticals, attainability and quality of data. The value of input parameters and assumptions has a direct correlation with the result. Which means the results will vary based on these two parameters. It is known that in ML, the more datasets and features we have to test and train, the outcome of the results will be better. One variable here would be that the possibility of the virus behaving differently over time is very high, as it mutates and adapts to different environments.

1.1. When is a patient said to be recovered?

A patient is said to be medically recovered from COVID-19 only after the antibodies that are fighting the infection have suppressed the virus successfully, preventing the virus from mutating and replicating and supporting the recovery process. When no long-term health effects or disabilities are found, the patient is said to be fully recovered. We use Machine Learning in our project to successfully predict the number of people recovering in a day to day basis. In this work, we try to conduct a thorough and precise prediction for the healthcare workers in the frontlines to be able to anticipate the recovered patients. This can be an extremely great boost especially for the morale of the general public, and our frontline COVID warriors.

Initially we set out to use Decision tree algorithm for the effective implementation of our project. But after further investigation and trails, we switched to Random forest Algorithm. The reasons for our switch are as follows;

1.2. Problems with Decision Trees

Although decision tree algorithm is an well-structured regression model, unfortunately there are still a couple of discrepancies which will obstruct the smooth implementation of decision trees Algorithm.

Some of the drawbacks are:

●A small change in data may cause a completely different set of information, thus causing the model to offer incorrect predictions.

●Decisions tree algorithms are extremely sensitive to the information they're trained on and even minute changes to the training set may result in vastly different tree structure.

●Decision trees tend to seek out locally optimal solutions instead of considering the globally optimal ones.

Such are the drawbacks of using Decision tree algorithm. So in order to overcome all the above mentioned problems, we went ahead with random forest.

2. Related Work/Literature Review

After going through a lot of other projects and published papers, To our knowledge we found no other project/paper that similarly emulates the same topic with the same implementation as ours.

The Already existing methods of such project which also use machine learning / Deep learning which come under the cover of COVID-19 are mainly focused on Prediction of general COVID-19 cases over a cumulative period of time. Alternatively they are focused on the prediction of active cases.

Our base paper concentrates on predictions of cases and recovery by using deep learning based nested sequence along with Long-short term memory Architecture. Their comprised dataset although very diverse in nature is small. And the prediction of their recovered cases is based on the prediction of their infection cases.

Similarly, some other projects from which we have taken inspiration and ideas include a project that uses logistic growth curve model for short term predictions and SIR models to forecast the maximum number of active cases and peak time; and Time Interrupted Regression model to evaluate the impact of lockdown and other interventions.

Here Attributes such as positivity ratio, population density and active cases were actually derived using other attributes.

The conclusion of their project is that it precisely predicts for a short duration in the case of India and other high incidence states. But since the mid/long term prediction is unknown, because of the mere fact of shortage of data availability, at their time of conducting their respective experiment.

The next project would be Using Artificial Intelligence Approach in order to predict the recovery of the patient who had been affected by the COVID-19 virus.

This project uses the artificial intelligence approach, using recurrent neural networks and support vector machine models.Most of the people who got infected who had cough, fever or fatigue couldn’t recover. Also, Patients living in or visited Wuhan, Iran, France, Italy were at a higher risk of dying.

Some of their demerits include, that more symptoms are being added as virus mutates which questions the feasibility and also the availability of data in their datasets. Since their study took place around March-April 2020, around the time when the pandemic was taking off, the probability of them having insufficient data was also very severe.

3. Data-set characteristics and Analysis

The attributes in our complete data set are as follows:

- Confirmed cases
- Deceased cases
- Tested cases
- Daily confirmed
- Daily deceased
- Daily recovered
- Active cases
- Positivity ratio
- Area (km^2)
- Total population (in 000)
- Population density/km^2
- States Encoded
- Day
- Month
- Year

Positivity ratio can be defined as the relationship between total tested positive patients with their net cases.

Active cases can be defined as the patients who currently undergoing the treatment process for COVID-19. In arithmetic terms, it can be defined as confirmed cases – recovered cases-deaths.

Population Density can be defined as the total number of people in a specific region divided by its area.

After all this, we then analysed all the compiled data which we procured from various sources, some of our sources include www.covid19india.org, www.kaggle.com, www.censusindia.gov.in.

Some examples of our compiled data is visualised in a bar graph format in the below provided diagram.

Fig 1 represents the cumulative confirmed cases for all states.

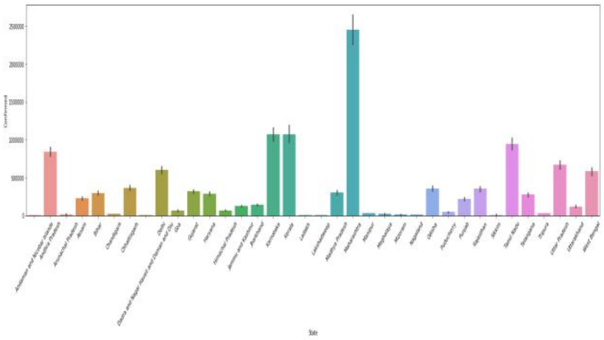


Fig 1.

The bar graph below represents the cumulative deceased cases for all states (fig 2).

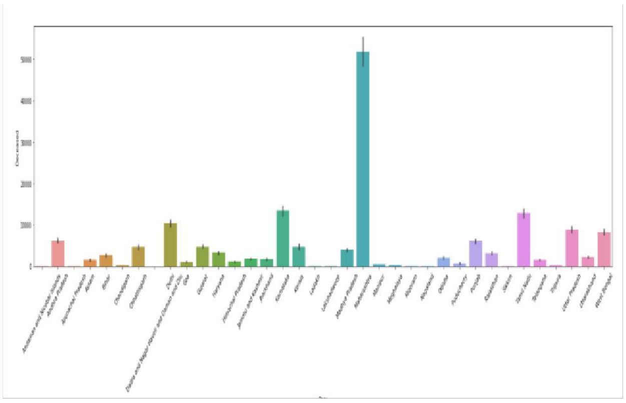


Fig 2.

4. Formula and Equations

Some of the attributes in our dataset like Population density, Positivity ratio and active cases are derived using the following formulas;

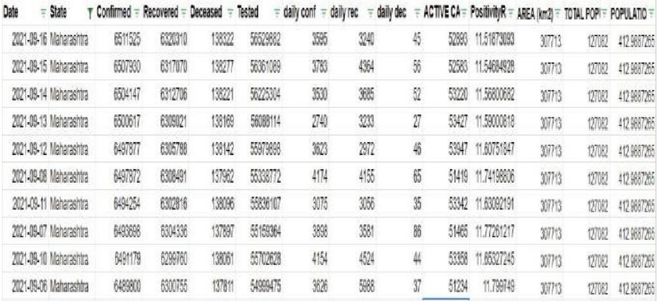
- Positivity Ratio = Confirmedcases (Tested Positive)/Total cases
- Active cases = confirmed cases-recovered cases- deaths
- Population Density = Total population/area

5. Methodology

We decided to go forward with Random forest regression technique as the algorithm to implement our project.

Random Forest Regression is said to be a supervised learning algorithm that employs a method called ensemble learning method for regression. Ensemble learning method is used to combine predictions from a variety of ML algorithms to assemble a more precise prediction than a single model in orderto vastly improve the accuracy of our predicted data. Initially, we gathered datasets from multiplesources, some are :

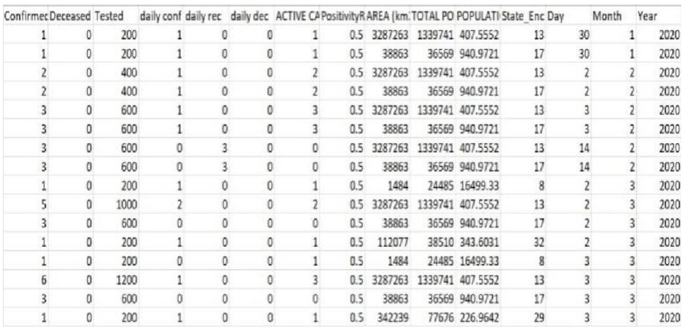
Our dataset consists around 18 months of daily data for all our attributes such as confirmed cases, deceased cases, etc. This dataset also comprises of data for all 36 states and union territories of India. Fig 3 represents an example of what our raw dataset looked like



Date	State	Confirmed	Recovered	Deceased	Tested	daily conf	daily rec	daily dec	ACTIVE CA	PositivityR	AREA (km2)	TOTAL POPI	POPULATIO
2021-09-16	Maharashtra	6511525	6320310	138322	56529882	3595	3240	45	52893	11.51873093	307713	127082	412.9887265
2021-09-15	Maharashtra	6507930	6317070	138277	56361089	3783	4364	56	52583	11.54684928	307713	127082	412.9887265
2021-09-14	Maharashtra	6504147	6312706	138221	56225304	3530	3685	52	53220	11.56800682	307713	127082	412.9887265
2021-09-13	Maharashtra	6500617	6309021	138169	56088114	2740	3233	27	53427	11.59000818	307713	127082	412.9887265
2021-09-12	Maharashtra	6497877	6305788	138142	55979898	3623	2972	46	53947	11.60751847	307713	127082	412.9887265
2021-09-08	Maharashtra	6497872	6308491	137962	55338772	4174	4155	65	51419	11.74198806	307713	127082	412.9887265
2021-09-11	Maharashtra	6494254	6302816	138096	55836107	3075	3056	35	53342	11.63092191	307713	127082	412.9887265
2021-09-07	Maharashtra	6493698	6304336	137897	55159364	3898	3581	86	51465	11.77261217	307713	127082	412.9887265
2021-09-10	Maharashtra	6491179	6299760	138061	55702628	4154	4524	44	53358	11.65327245	307713	127082	412.9887265
2021-09-06	Maharashtra	6489800	6300755	137811	54999475	3626	5988	37	51234	11.799749	307713	127082	412.9887265

Fig 3.

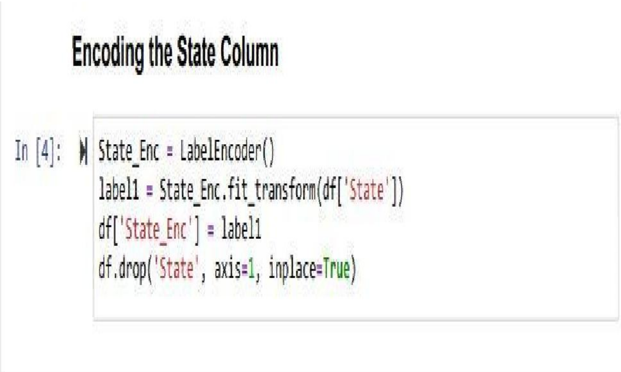
Data pre-processing is the next stage, this data was first arranged in ascending order, and then this pre-processed data was split into 2 parts. Where one part is used for training and the other part is used for testing. You can see an example of the processed data in fig 4.



Confirmec	Deceased	Tested	daily conf	daily rec	daily dec	ACTIVE CA	PositivityR	AREA (km:	TOTAL PO	POPULATI	State_Enc	Day	Month	Year
1	0	200	1	0	0	1	0.5	3287263	1339741	407.5552	13	30	1	2020
1	0	200	1	0	0	1	0.5	38863	36569	940.9721	17	30	1	2020
2	0	400	1	0	0	2	0.5	3287263	1339741	407.5552	13	2	2	2020
2	0	400	1	0	0	2	0.5	38863	36569	940.9721	17	2	2	2020
3	0	600	1	0	0	3	0.5	3287263	1339741	407.5552	13	3	2	2020
3	0	600	1	0	0	3	0.5	38863	36569	940.9721	17	3	2	2020
3	0	600	0	3	0	0	0.5	3287263	1339741	407.5552	13	14	2	2020
3	0	600	0	3	0	0	0.5	38863	36569	940.9721	17	14	2	2020
1	0	200	1	0	0	1	0.5	1484	24485	16499.33	8	2	3	2020
5	0	1000	2	0	0	2	0.5	3287263	1339741	407.5552	13	2	3	2020
3	0	600	0	0	0	0	0.5	38863	36569	940.9721	17	2	3	2020
1	0	200	1	0	0	1	0.5	112077	38510	343.6031	32	2	3	2020
1	0	200	0	0	0	1	0.5	1484	24485	16499.33	8	3	3	2020
6	0	1200	1	0	0	3	0.5	3287263	1339741	407.5552	13	3	3	2020
3	0	600	0	0	0	0	0.5	38863	36569	940.9721	17	3	3	2020
1	0	200	1	0	0	1	0.5	342239	77676	226.9642	29	3	3	2020

Fig 4.

Other processes that we did in our dataset include “Encoding the state column”, which had to be done as the data present in our dataset vary over a period of 18 months, where the first six months are repeated after 12 months, which causes an error. Hence the change had to be made. Shown in Fig 5.

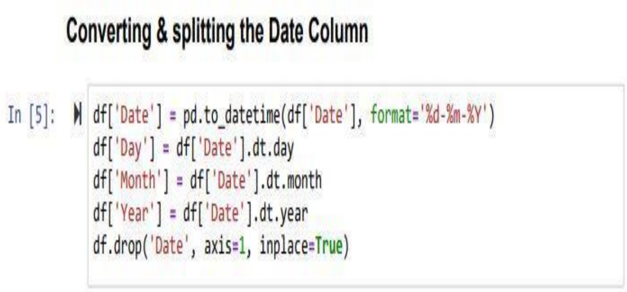


Encoding the State Column

```
In [4]:  State_Enc = LabelEncoder()
         label1 = State_Enc.fit_transform(df['State'])
         df['State_Enc'] = label1
         df.drop('State', axis=1, inplace=True)
```

Fig 5.

Followed by converting and splitting the date column. This is represented in Fig 6.

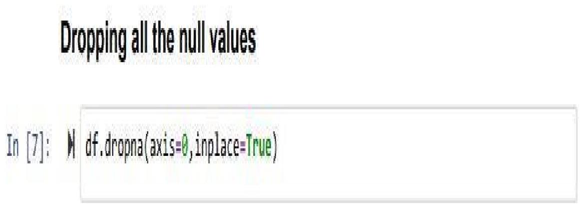


Converting & splitting the Date Column

```
In [5]:  df['Date'] = pd.to_datetime(df['Date'], format='%d-%m-%Y')
         df['Day'] = df['Date'].dt.day
         df['Month'] = df['Date'].dt.month
         df['Year'] = df['Date'].dt.year
         df.drop('Date', axis=1, inplace=True)
```

Fig 6.

Followed by dropping all null values, This is represented in Fig 7.



Dropping all the null values

```
In [7]:  df.dropna(axis=0,inplace=True)
```

Fig 7.

We have used Python programming language to implement and execute this project. Since this project falls under the scope of Machine Learning, we had to make use of some necessary libraries. Some of theimportant libraries used to facilitate this project are as follows;

- Pandas

Pandas is an open-source, BSD-approved Python library, giving high execution, easy to-use data designs and data assessment devices for the Python programming language.

This library is primarily built on top of another library called numpy. Python with Pandas is used in a wide extent of fields including academic and business spaces which incorporate money, monetary angles, Statistics, assessment, and many more

- NumPy

NumPy is a python library which actually stands for Numerical Python, it is a library involving multidimensional display objects and parcel of choices of schedules for dealing with those clusters. Using NumPy, mathematical and intelligent procedures can be performed on exhibits.

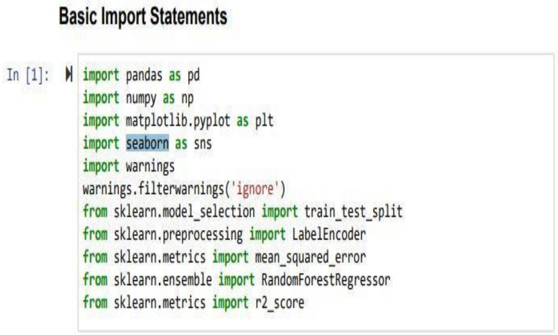
- matplotlib.pyplot

matplotlib.pyplot is a data visualization, cross platform, graph plotting library used in python. It provides a MATLAB-like method of plotting. The central use of pyplot is for making interactive plots.

- Seaborn

Seaborn is an information representation library, which is based on top of matplotlib and firmly incorporated with pandas information structures in python.

Representation is the main point of Seaborn, which helps in investigation and apprehension of information. The above mentioned libraries can be seen in fig 8.



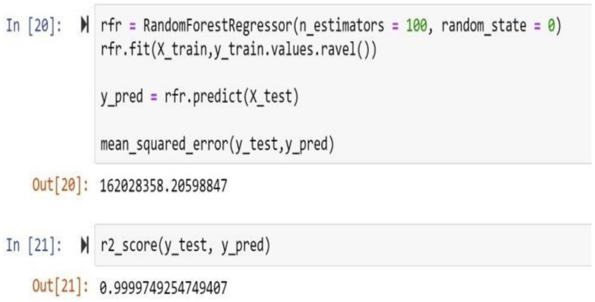
Basic Import Statements

```
In [1]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns
         import warnings
         warnings.filterwarnings('ignore')
         from sklearn.model_selection import train_test_split
         from sklearn.preprocessing import LabelEncoder
         from sklearn.metrics import mean_squared_error
         from sklearn.ensemble import RandomForestRegressor
         from sklearn.metrics import r2_score
```

Fig 8.

We then execute all the statements. We used jupyter notebook as our preferred application to implement and execute our project.

6. Results and Discussions



```
In [20]:  rfr = RandomForestRegressor(n_estimators = 100, random_state = 0)
          rfr.fit(X_train,y_train.values.ravel())

          y_pred = rfr.predict(X_test)

          mean_squared_error(y_test,y_pred)

Out[20]:  162028358.20598847

In [21]:  r2_score(y_test, y_pred)

Out[21]:  0.9999749254749407
```

Fig 9.

As we can see from the above fig 9,we conclude that the results of this project are fairly promising. With an r2 value of 0.99 this estimation can be deemed as strong effect size. The range of values for R-Squared can go from 0 to 1. Hence we can concur that the value 0 shows that the response variable can't be explained by the indicator variable by any means. Whereas a value of 1 shows that the reaction variable can be impeccably clarified without mistake by the indicator variable. In Real time practice, the likelihood of getting the value 0 or 1 is practically none.

The mean squared error of this project is 162028358.20598847, which when calculated and rounded off to the nearest lower end will give us around 12,000.

Below you can see a bar graph (Fig 10) in which we visually represent the actual recovered cases and our projects predicted recovered cases, as we can see the discrepancy is fairly small.

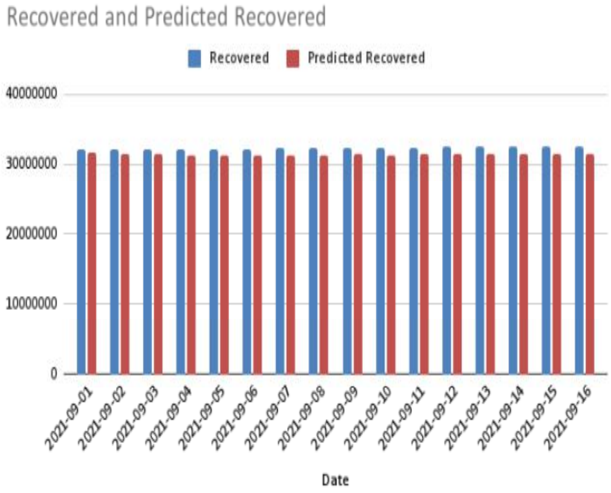


Fig 10.

7. Scope and Future Works

The scope of this project is limitless and can be used and referred to in many other significant studies/projects in this respective domain. This Project can also be referred to by government officials to get an brief understanding of the amount of patients that are expected to recover on a daily basis.

That will in turn be tremendously helpful is analyzing or predicting the recovery of the country's economy. It can also give us a brief understanding of hospital beds that are going to be potentially free. And give a much needed mental break for our frontline warriors.

This project would also be greatly beneficial to the medical industry. From calculating the beds that are going to be left open so others can get the treatment needed, to many others.

This project would also be greatly beneficial to other researchers to test out their theories, such as herd immunity etc.

Finer tuning of the project has to be done. The Promising results at this stage show us that the forthcoming results will definitely not disappoint.

8. Conflicts of Interest

The authors whose names are listed attest that they bear NO coalition or involvement in any association or entity with any fiscal interest or financial interest in the subject matter or materials discussed in this project. The authors bear no conflict of interest while working on this project.

9. Acknowledgments

We would like to thank our corresponding author Dr A. Pandian for his mentorship and continued support throughout this project. We also would like to thank SRM University for providing a platform in which we could receive all necessary guidance's and mentorship.

10. Author Contributions

Conceptualization was done by Adhithya and Raghu; methodology was done by Adhithya and Raghu; Software used was decided by Raghu and Adhithya; The Validation was performed by Adhithya, Raghu and Dr. Pandian; Formal analysis was conducted by Raghu and Adhithya; Investigation was conducted by Adhithya and Raghu; Resources were collected by Raghu b; Data curatio was done by Raghu b; Writing—original draft preparation was done by Adhithya; Writing—review and editing was carried out byAdhithya and Raghu; Visualization was performed byAdhithya and Raghu ; Supervision was done by Dr A Pandian; Project administration was done by Dr A Pandian; .

11. References

1] Li Q, Guan X, Wu P, et al. Early transmission dynamics in wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med. 2020;382(13):1199–1207.

2] WHO. Novel Coronavirus – China. [Internet]. Geneva: World Health Organization; 2020. Available from: https://www.who.int/csr/don/12-january-2020-novelcoronavirus-china/en/.

3] JHU-CSSE. COVID-19 Dashboard. [Internet]. Baltimore, Maryland: Center for Systems Science and Engineering (CSSE) at Johns Hopkins University; 2020. Available from: https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6.

4] Cohen J, Kupferschmidt K. Mass testing, school closings, lockdowns: countries pick tactics in 'war' against coronavirus. Science. . [Internet], Available from:. https://www.sciencemag.org/news/2020/03/mass-testing-school-closings-lockdownscountries-pick-tactics-war-against-coronavirus#; 2020.

5] Zhang Y, Jiang B, Yuan J, Tao Y. The Impact of Social Distancing and Epicenter Lockdown on the COVID-19 Epidemic in Mainland China: A Data-Driven SEIQR Model Study. 2020; 2020.

6] MoHFW. COVID-19 Statewise Status. [Internet]. New Delhi: Ministry of Health and Family Welfare, Government of India; 2020. Available from: https://www.mohfw.gov.in/.

7] Batista, M. (2020). Estimation of the final size of the COVID-19 epidemic. Preprint.] medRxiv.

8] S. Bhatnagar, V. Lal, S.D. Gupta, O.P. Gupta, Forecasting incidence of dengue in Rajasthan, using time series analyses, Indian J. Public Health 56 (4) (2012) 281.

9] M. Widerström, M. Omberg, M. Ferm, A.K. Pettersson, M.R. Eriksson, I. Eckerdal, J. Wiström, Autoregressive integrated moving average (ARIMA) modeling of time series of local telephone triage data for syndromic surveillance, Online J. Public Health Inform. 6 (1) (2014).

10] M. Manikandan, A. Velavan, Z. Singh, Forecasting the trend in cases of Ebola virus disease in west African countries using auto regressive integrated moving average models, Int. J. Commun. Med. Public Health 3 (2016) 615–618.

11] Ranjan R. Predictions for COVID-19 outbreak in India using Epidemiological models. medRxiv. 2020 2020.04.02.20051466.

12] Ray D, Salvatore M, Bhattacharyya R, et al. Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: data science call to arms. medRxiv. 2020 2020.04.15.20067256.

13] Bhatnagar M. COVID-19: Mathematical Modeling and Predictions. 2020; 2020.

14] Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. Springer Nature Switzerland AG 2020:Modeling Earth Systems and Environment (2021) 7:1385–1391

15] Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India ELSEVIER:Chaos, Solitons and Fractals 139 (2020) 110017

16] Who is dying from COVID-19 and when? An Analysis of fatalities in TamilNadu, India ELSEVIERClinical Epidemiology and Global health 9(2021):275-279

17] When COVID-19 will decline in India? Prediction by combination of recovery and case load rate ELSEVIER Clinical Epidemiology and Global Health 9 (2021) 17–20

18] ARIMA modeling & forecasting of COVID-19 in top five affected countries Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models ELSEVIER: Clinical Epidemiology and Global health 9(2021) 26-33

19] Prediction for the spread of COVID-19 in India and effectiveness of preventive measure ELSEVIER: Science of total environment 728(2020) 138762

20] Public perception and preparedness for the pandemic COVID A Health Belief Model approach ELSEVIER: Clinical Epidemiology and Global Health 9 (2021) 41–46

21] T. Kuniya, "Prediction of the Epidemic Peak of Coronavirus Disease in Japan 2020", Journal of Clinical Medecine, 9:789, doi:10.3390, 2020

22] S. Jung, A.R. Akhmetzhanov, K. Hayashi, N.M. Linton, Y. Yang, B. Yuan, T. Kobayashi, R. Kinoshita and H. Nishiura, “RealTime Estimation of the Risk of Death from Novel Coronavirus (COVID-19) Infection: Inference Using Exported Cases”, Journal of Clinical Medecine, 9:523, doi:10.3390/jcm9020523, 2020.

23] J. Brownlee, “Long Short-Term Memory Networks With Python, Machine Learning” Mastery Edition, 2017

24] “Epidemiological data from the COVID-19 outbreak, real-time case information”, BoXu, BernardoGutierrez, Sumiko Mekaru,, Kara Sewalk, LaurenGoodwin, Alyssa Loskill, Emily L.Cohn, Yulin Hswen, SarahC. Hill , Maria M. Cob, Alexander E. Zarebsk, Sabrina Li, Chieh-HsiWu, Erin, Julia D., , KatelynnO’Brien, SamuelV. Scarpino8, John S. Brownstein,, OliverG. Pybus, David M. Pigott, Moritz U.G. Kraemer.

25] Wu JT, Leung K, Leung GM, Now casting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modeling study. Lancet,2020, In pressing. DOI: 10.1016 / S0140- 6736 (20) 30260-9