A Study On Covid-19 Data Of India, Andhra Pradesh And Telangana Using Machine Learning Algorithms

[1]K.L.S. Soujanya, Challa Madhavi Latha[2], [3]N. Sandeep Chaitanya

[1]Professor, Dept. of Computer Science & Engineering, CMR College of Engineering & Technology, Kandlakoya, Hyderabad, India.
[2]Assistant Professor, Dept. of Computer Science & Engineering, CMR College of Engineering & Technology, Kandlakoya, Hyderabad, India.
[3]Assistant Professor, Dept. of Computer Science & Engineering, VNRVJIET, Hyderabad, India.

Email: [1]souj47@gmail.com, [2]saidatta2009@gmail.com, [3]sandeepchaitanya_n@vnrvjiet.in

Abstract: The epidemic of Covid-19 has created a disastrous situation around the globe. The spread of Covid-19 is drastically increasing day by day. Machine learning is one of the efficient tools to track the outbreak of the disease, forecast the probable confirmed and death cases as well as the fatality rate. This study applies multiple regression analysis which is one of the supervised machine learning algorithms to analyze and forecast the fatality rate. The study was conducted to predict the spread of Covid-19 in areas of Telangana, Andhra Pradesh, and India. R-Square (R2), Mean square error (MSE), Root mean square error (RSME) and Mean absolute error (MAE) are the main measures used to predict the accuracy of the algorithm. The results reveal that the case fatality rate is higher in Telangana compared to Andhra Pradesh and India, and more diseased cases are observed in Andhra Pradesh. The study was conducted with the available data; if sufficient data is available then the more precise predictions could be possible using multiple regression analysis.

Keywords: Regression analysis, COVID-19, Mean square error, Mean absolute error, Machine learning algorithms, Disease.

1. INTRODUCTION

Machine learning (ML) has raised popularity by proving itself over a decade as it is solving the real-time complex problems. ML real-time applications are incorporated in maximum domains such as e-health, e-business, robotics, gaming, weather predictions, language, voice, image processing, etc. ML is the controversy of conventional programming because ML follows trial and error methods whereas conventional programming follows conditional statements. ML can predict the future actions of numerous areas such as disease forecasting (Wu et al. 2020), weather forecasting, stock price forecasting, etc. In the modern era, ML algorithms have more potential techniques to improve the accuracy of forecasting. The significant techniques for the prediction in ML are support vector machines, linear regression, logistic regression, naive Bayes, decision trees, K- nearest, neural networks, random forest, gradient boosting, etc. However, there is a need to develop novel techniques for forecasting to avoid overfitting and underfitting levels. The Healthcare sector is also using the ML techniques to predict the disease and predict the situation of patients. Many studies have been carried out to predict the various diseases such as severe acute respiratory syndrome SARs, cardiovascular, diabetics, cancer, heart issues, etc.(Guan et al. 2020; Song et al. 2020; Yin & Wonderlink, 2018; de et al. 2016; Zumla et al. 2016; Drosten et al. 2003; Guan 2003).

The main aim of the paper is to forecast the case fatality rate (CFR) of Novel Coronavirus disease 2019 (COVID-19) by using confirmed cases of Telangana, Andhra Pradesh, and India. COVID-19 was identified in December 2019, in Wuhan, China, and spread over the world very quickly (Huang et al. 2020; Chen et al. 2020). The world is frightening by COVID-19 and it becomes a global pandemic, which was declared by world health organization (WHO) (Lu et al. 2020; Li et al. 2020). In India, the first case of COVID-19 was reported on January 30th, 2020 in Kerala state for a student, who came from Wuhan city, China. In India, the mass transmission of the disease was detected after Tablighi Jamaat religious congregation in Delhi's Nizamuddin area in March 2020.

The union health ministry of India has declared that the more cases of COVID-19 have been found in almost all states in India because of the Markaz event. Joint Secretary in the health ministry Lav Agarwal confirmed that the most of the Markaz related cases are leading to high risk in various states, such as Tamil Nadu (84%), Telangana (79%), Delhi (63%), Utter Pradesh (59%) and Andhra Pradesh (61%). According to the data from the Indian council for medical research (ICMR) https://www.covid19india.org/, updated till 18th June of 2020, India has reported 368648 confirmed cases with COVID-19, death cases are 12275; 32% of highest cases identified in Maharashtra state. Moreover, globally 213 countries were reported on 18th June 2020, a total confirmed cases 8428263, deaths recorded as 451916 cases were reported by WHO. The top 5 most infected countries are the United States of America with 26.51% (2234854), Brazil with 11.39% (960309), Russia with 6.65% (561019), India with 4.37% (368648), the United Kingdom with 3.55% (299251) cases.

Melin et al. (2020) analyzed the COVID19 pandemic around the world using self-organizing maps. They concluded that the clustering is possible to group countries based on COVID confirmed cases. Several studies on deep learning have been reported significant accuracy for auto-detection of various diseases such as lung diseases and skin diseases based on images (Ardila et al. 2019; Suzuki, 2017; Coudray et al. 2018; He et al. 2015; Esteva et al., 2017). Many studies on regression and neural network models have been predicting the patients' condition for a specific disease, for instance, coronary artery disease, cardiovascular disease, breast cancer and COVID-19 forecasting (Harrell et al. 1995; Lapuerta et al. 1995; Anderson et al. 1991; Asri et al. 2016; Petropoulos and Makridakis 2020). The forecasting of disease is very useful to make decisions to handle the situation and guide to manage diseases effectively.

ML algorithms are very useful for COVID-19 prediction to deal with the current scenario and provide the precautions to manage disease very efficiently. The main objective of the present study is to forecast the spread of COVID-19 in Telangana state, Andhra Pradesh, and India. Thousands of people are highly infected by this pandemic throughout India, the focus of the study in this paper is Telangana and Andhra Pradesh states. Daily hundreds of new cases have been reported as well as recovered cases, and few of death cases in India. The main reasons for spreading viruses are close contact with persons, touching the contaminated surfaces, and infected droplets. A person can spread the virus without knowingly as COVID-19 is an asymptotic virus. Worldwide medical and drug researchers have been involved to invent the vaccine and medications to cure the patients. However, till now this research is constrained to the clinical lab trails and the medications or vaccine not yet discovered. Hence, it is declared a dangerous virus, so almost the entire globe has been following lockdowns to stop the virus spread in affected areas.

Telangana and Andhra Pradesh states are not exempted from this; strict lockdown had followed after a 14-hour voluntary public curfew on 22nd March, on 24th March the Government of India declared 21 days nationwide lockdown to prevent the spread of COVID-19 in India. The Phase-I lockdown period starts from March 25th, 2020 to April 14th, 2020. Before reaching the end of the Phase-I period many state governments decided to extend the lockdown till the end of April, among the states Telangana state is one of the states. Phase-II lockdown period is April 15th, 2020 to May 3rd, 2020. In this period the areas of lockdown have been divided into 3 zones, one in the red zone, which indicates the high infected area, the second one is orange zone specifying few COVID-19 cases area, the last one is green zone represents without any infections. April 19th, 2020 onwards, the partial lockdown was followed in Telangana. Phase- III starts from May 4th, 2020 to May 17th, 2020, Phase –IV period from May 18th, 2020 to May 31st, 2020, and these two periods are considered as partial lockdown periods. Due to focus on economic growth, Phase – V termed as unlock-I period, except large gatherings, everything has been reopened. After all these precautions also COVID-19 has not yet stopped, it is spreading numerously. The present paper is contributing to this aspect of information and studies the various dimensionalities to help the people to reduce the crisis. This paper attempts to analyze the COVID-19 situation of Telangana, Andhra Pradesh states, and India in the phase-wise and cumulative study up to June 30th, 2020.

2. MATERIALS AND METHODS

Dataset:

The main aim of this paper is to analyze and forecast the COVID-19 spread over the Telangana, Andhra Pradesh, and India. It is focusing on new infected cases, deaths, and discharged patients. The dataset was collected from March 25th, 2020 to June 30th, 2020 from the ICMR and WHO websites.

Analysis:

The present study is analyzed based on the following three categories in Telangana.

1. Before lockdown
2. Lockdown phases (Phase1 to Phase 4)
3. Unlock 1

Table1: Phase wise infected cases in Telangana (25th March to 30th June)

Phase Num	Hospitalized	Recovered	Diseased
Phase-0(upto 24th March)	20	0	0
Phase-1	605	0	1
Phase-2	438	229	4
Phase-3	469	447	5
Phase-4	1147	436	48
Phase-5	13641	5866	178

Source: Compiled by authors

Before lockdown:

Telangana state is not native land of COVID-19, even though people were affected due to close contact with COVID patients. Most of the people were infected by COVID-19 because of the following two reasons. One is foreign nationals and the other one is traveling from other countries. The following Table 2 is depicting the information of COVID patients hospitalized before lockdown. Table 2 is showing that most of the people from the Telangana state affected in the Hyderabad district only and very few from Karimnagar district. Moreover, people were affected by Indonesians and traveled from other countries.

Table2: People infected by localities or foreigners.

Date Announced	Detected District	Current Status	Nationality	Type of transmission
2/3/2020	Hyderabad	Recovered	India	Imported
14/03/2020	Hyderabad	Hospitalized	India	Imported
15/03/2020	Hyderabad	Hospitalized	India	Imported
16/03/2020	Hyderabad	Hospitalized	India	Imported
17/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
18/03/2020	Hyderabad	Hospitalized	India	Imported
18/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
18/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
18/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
18/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
18/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
18/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
18/03/2020	Hyderabad	Hospitalized	Indonesia	Imported
19/03/2020	Hyderabad	Hospitalized	India	Imported
19/03/2020	Hyderabad	Hospitalized	India	Imported
19/03/2020	Hyderabad	Hospitalized	India	Imported
20/03/2020	Hyderabad	Hospitalized	India	Imported
20/03/2020	Karimnagar	Hospitalized	Indonesia	Imported
20/03/2020	Karimnagar	Hospitalized	Indonesia	Imported
21/03/2020	Hyderabad	Hospitalized	India	Local
21/03/2020	Hyderabad	Hospitalized	India	Imported

Lockdown phases (Phase 1 to Phase 4):

The government of India has proposed lockdown to mitigate the spread of COVID-19. The lockdown was implemented in 4 phases which were started from March 25th, 2020 to May 31st, 2020. Table 1 shows the phase-wise infected cases in Telangana state during this period. It is observed that the spreading of COVID-19 in the state of Telangana was very slow and very less when compared to phase 5. In the entire state of Telangana, 0.0075 percent of the population was affected by the decease and 2.1813 percent of the affected have died.

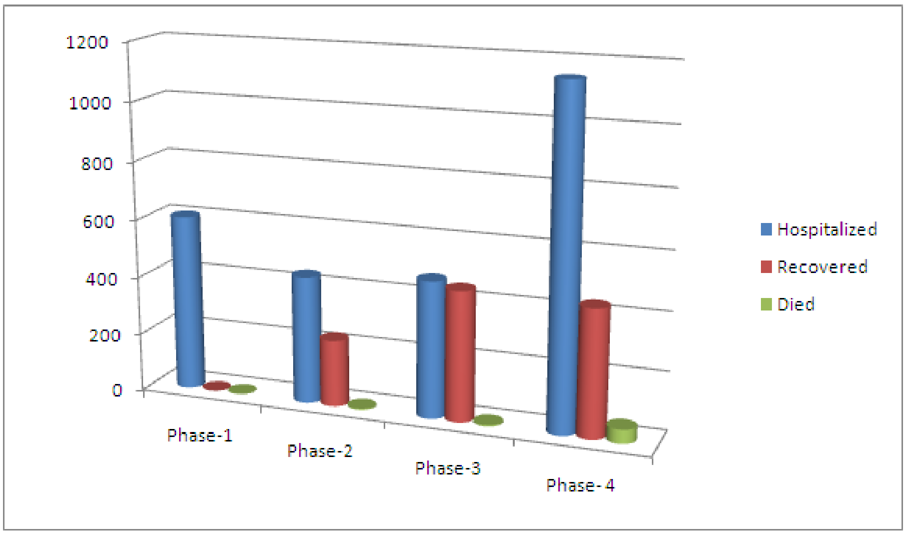

Figure 1: Covid-19 data of Phase-1 to Phase -4 for Telangana state

Phase 5 or Unlock-1 phase:

In the view of the downfall of the economy and survival of daily wages labor the government of India has lifted the lockdown partially named as an unlock-1 phase. Though the government insisted to follow the precautions, there was a huge rise in the spread of disease. In the state of Telangana, 0.0388 percent of the population was affected by COVID-19 and 1.305 percent of affected people have died.

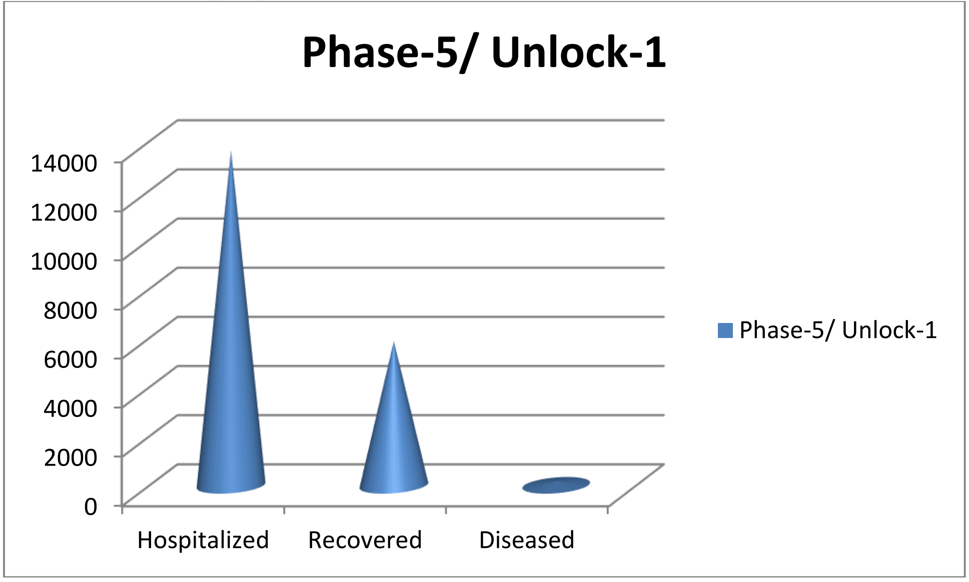

Figure 2: Phase 5 cases for Telangana state.

Area –wise Analysis:

In Telangana state there are 33 districts, almost all districts are affected by Covid-19. Among affected lot 80 percent are from Hyderabad and the remaining 20 percent are from other districts. The death rate is also high in Hyderabad which is 76 percent of death cases in Telangana. Very few cases have been recorded in Narayanpet, Peddapalli, and Wanaparthy. The flowing Table 3 shows the descriptive statistics of district-wise information under Telangana state. The average of confirmed cases is 503.58, active cases are 485.97, Recovered 16.7, deceased 0.91. In the view of average cases very less recorded for deceased and recovered cases. The highest Standard deviation shows for the confirmed and active cases, which means affected people are more than the recovered and deceased cases.

Table 3: Descriptive statistics for district-wise data

	Confirmed	Active	Recovered	Deceased	CFR
Count	33.00	33.00	33.00	33.00	33.00
Mean	503.58	485.97	16.70	0.91	1.19
Std	2328.59	2272.87	52.46	4.00	5.80
Min	3.00	2.00	0.00	0.00	0.00
25%	24.00	21.00	1.00	0.00	0.00
50%	47.00	32.00	4.00	0.00	0.00
75%	82.00	70.00	15.00	0.00	0.00
Max	13422.00	13094.00	305.00	23.00	33.33

Source: Compiled by authors

Correlation:

Table 4 depicts the correlation analysis of Telangana state district-wise. The output variable CFR is positively significant on deceased cases, deceased cases are significantly dependent on confirmed and active cases. Recovered cases are equally significant in confirmed, active and deceased cases. There is a strong correlation between the deceased and the CFR.

Table 4: Correlation analysis

	Confirmed	**Active**	**Recovered**	**Deceased**	**CFR**
Confirmed	1				
Active	0.999893	1			
Recovered	0.986529	0.984038	1		
Deceased	0.9963617	0.993139	0.9836411	1	
CFR	-0.039009	-0. 038748	-0.053477	0.010165	1

Source: Compiled by authors

Note: CFR: case fatality rates

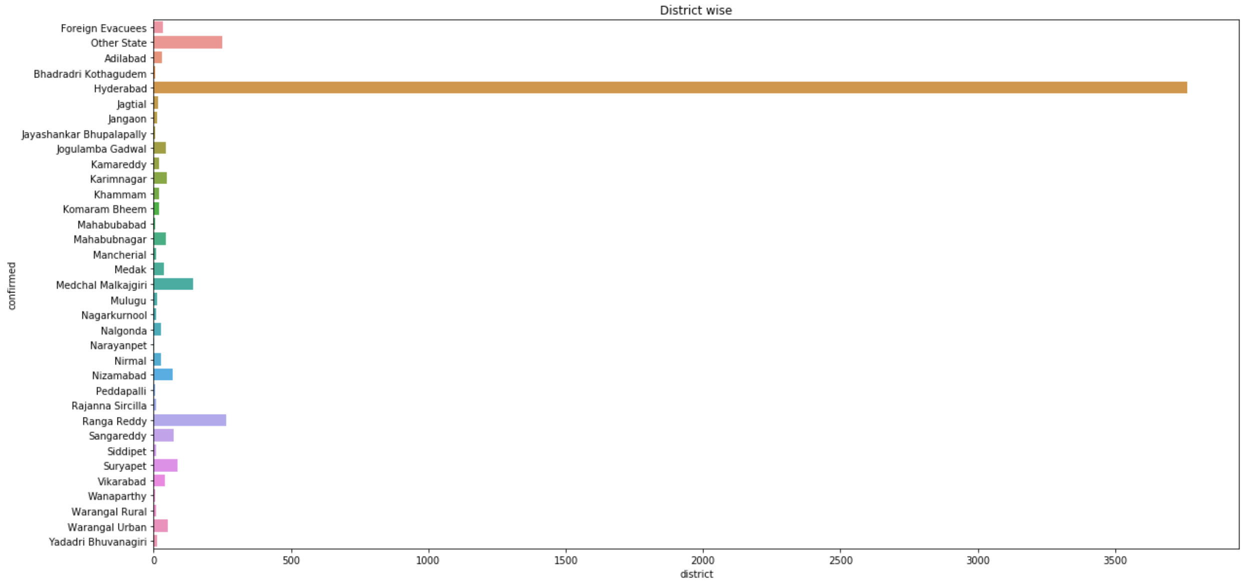

Figure 3: District wise detected cases:

Figure 3 shows the district wise detected cases in the Telangana region. It is observed that the highest detected cases are in Hyderabad followed by Ranga Reddy and Medchal districts.

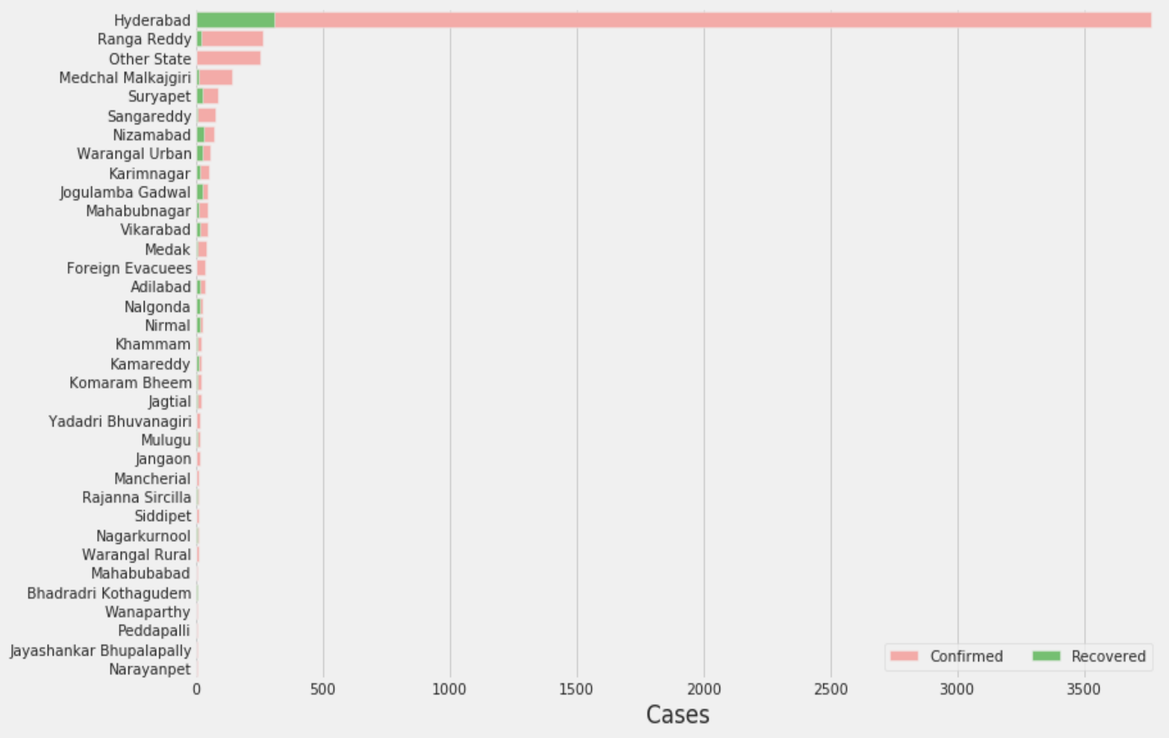

Figure 4: District wise confirmed and recovered cases

Figure 4 depicts the confirmed and recovered cases of district wise. It is observed that the recovered cases are very less when compare to confirmed cases as on June 30th, 2020.

Comparative analysis of Telangana, Andhra Pradesh, and India Covid-19 data

The present study on Covid-19 has been compared with two states and country information. The comparative analysis has been done by using multiple regression of Telangana, Andhra Pradesh, and India information. The extension of linear regression is multiple linear regression, which is used to forecast the value of the dependent variable on explanatory variables. In this study, CFR is a dependent variable, and the explanatory variables are confirmed, death, active and recovered cases. The output of the multiple regression analysis is shown in Table 5. Model accuracy and model fitting can be identified by the various parameters such as intercept, coefficients, train score, test score, R-Square (R2), Mean square error (MSE), Root mean square error (RSME) and Mean absolute error (MAE).

Table 5: Multiple Regression Analysis

State/ country	**Intercept**	**Coefficients**	**Train Score**	**Test Score**	**R^2**	**MSE**	**RMSE**	**MAE**
Telangana	2.7017	[0.00546504 -0.00564193 -0.00675266 0.01785963]	0.0304	-0.0272	0.00631	225.304	15.01	5.671
Andhra Prades	1.4370	[0.01286149 -0.01259072 -0.01462836 0.04008058]	0.0739	0.01167	0.072	3.80443	1.950	1.2498

h								
India	2.8297	[0.00019357 -0.0002254 -0.00021122 0.00063019]	0.0796	-0.01176	0.07040	2.2052	1.485	1.0974

Source: Compiled by authors

Table 5 depicts the result of multiple regression analysis of three regions i.e. Telangana, Andhra Pradesh, and India. Intercept shows the constant term which is also called bias, from this point the exploratory variables can start moving towards the upward direction. In the case of Telangana, 2.7 has been recorded which is almost nearer the intercept of India i.e. 2.83. It reveals that the highest CFR in the case of Telangana. Andhra Pradesh has been recorded as 1.44, which is less than the remaining two cases. The coefficient values represent the change of one unit of the dependent variable is explained by the independent variables. Coefficient indicates the positive or negative relationship between the dependent variable and its exploratory variables. Positive values represent the directly proportional relationship which specifies the linear dependent on exploratory variables, whereas negative values represent the inversely proportional relationship.

In the case of Andhra Pradesh, CFR is negatively significant at a 5 percent level of confidence with the recovery rate i.e. -0.014. In all the above cases the CFR is positively significant at the 10 percent level of confidence in the deceased category i.e. Telangana, Andhra Pradesh & India has recorded as 1.8 percent, 4 percent, and 0.6 percent respectively. This indicates the deceased are more in Andhra Pradesh. The R^2 value is the proportion of variance in the CFR variable on exploratory variables. In the case of Andhra Pradesh and India 7 percent of the R-square value, which means exploratory variables explain the 7 percent of the variability of the CFR variable. MSE, RMSE, and MAE variables are used to calculate the errors between forecasted and actual values. The present model shows less accuracy due to a lack of sufficient COVID-19 data.

3. CONCLUSION

The COVID-19 pandemic has shaken the world leading to a global crisis. The major population of the world is affected in terms of various aspects such as health, economy, education, social relationship, employment, transport sectors, etc. The present study is based on multiple regression analysis to predict the risk of COVID-19 in terms of CFR. This study analyzed day-wise historical data, performance, and accuracy. The results reveal that the highest CFR is found in Telangana state, it indicates the fatality rate is more when compared to Andhra Pradesh and India. CFR is positively significant at the 10 percent confidence level on the deceased category i.e. Telangana, Andhra Pradesh, and India. Artificial Intelligence and Machine Learning techniques can be used to analyze, understand, and predict the situations when sufficient data is available. The main limitation of the present study is insufficient datasets. However, the analysis of COVID-19 done as a part of the work is useful to understand the scenario of Telangana, Andhra Pradesh, and India. The present analysis is useful to make timely decisions to control the risk of COVID-19 by the respective authorities. The study will be extended to explore the prediction methodologies with updated datasets and use more accurate and appropriate ML algorithms. Real-time predictions can be done as a part of future work.

4. REFERENCES

[1] Huang. C et al., “Clinical features of patients infected with 2019 novel coronavirus in wuhan, china,” *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[2] Lu H, Stratton CW, Tang Y-W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. J Med Virol. 2020;92:401–402. 10.1002/jmv.25678

[3] Chen. N et al., “Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study,” *The Lancet*, 2020.

[4] “International committee on taxonomy of viruses (ictv) website,” https: //talk:ictvonline:org/, accessed 14 Feb 2020.

[5] “World health organization (who) website,” https://www:who:int/docs/default-source/coronaviruse/situation-reports/20200213-sitrep-24-covid-19:pdf?sfvrsn=9a7406a4 4, accessed 15 Feb 2020.

[6] Li. Q *et al.*, “Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia,” *New England Journal of Medicine*,2020.

[7] Wu. J.T, K. Leung, and G. M. Leung, “Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study,” *The Lancet*, 2020.

[8] “National health commission of the people’s republic of china website. 2020,” http://www:nhc:gov:cn/xcs/yqtb/202002/553ff43ca29d4fe88f3837d49d6b6ef1:shtml, accessed 14 Feb 2020.

[9] “Coronavirus disease (covid-19) situation dashboard,” https://www:who:int/redirect-pages/page/novel-coronavirus-(covid-19)- situation-dashboard, accessed 30 Mar 2020.

[10] “World health organization (who) website,” https://www:who:int/news-room/detail/30-01-2020-statement-on-the-second-meeting-ofthe-international-health-regulations-(2005)-mergency-committeeregarding-the-outbreak-of-novel-coronavirus-(2019-ncov), accessed 30 Mar 2020.

[11] “World health organization (who) website,” https://www:who:int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-mediabriefing-on-covid-19---28-february-2020, accessed 30 Mar 2020.

[12] Guan W.-j et al., “Clinical characteristics of 2019 novel coronavirus infection in china,” *MedRxiv*, 2020.

[13] Lei. J, Li. J, Li. X, and Qi. X, “Ct imaging of the 2019 novel coronavirus (2019-ncov) pneumonia,” *Radiology*, p. 200236, 2020.

[14] Song. F et al., “Emerging coronavirus 2019-ncov pneumonia,” *Radiology*, p. 200274, 2020.

[15] Chung. Met al., “Ct imaging features of 2019 novel coronavirus (2019-ncov),” *Radiology*, p. 200230, 2020.

[16] XINHUANET News Report. http://www.xinhuanet.com/english/2020-01/09/c_138690570.htm. Accessed February 6, 2020.

[17] Yin Y, Wonderlink RG. MERS, SARS and other coronaviruses as causes of pneumonia. Respirology. 2018;23:130-137.

[18] de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016;14:523-534.

[19] Zumla A, Chan JF, Azhar EI, Hui DS, Yuen KY. Coronaviruses–drug discovery and therapeutic options. Nat Rev Drug Discov. 2016;15:327-47.

[20] Drosten C, Günther S, Preiser W, et al. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med. 2003;348:1967-1976.

[21] Guan Y. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science. 2003;302:276-278.

[22] Cavanagh D. *Coronaviridae*: a review of coronaviruses and toroviruses In: Schmidt A, Wolff MH, Weber O, eds. Coronaviruses with Special Emphasis on First Insights Concerning SARS. Switzerland: Birkhauser Verlog Basel; 2005.

[23] Melin, J.C. Monica, D. Sanchez, O. Castillo, (2020), Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps, Chaos, Solitons & Fractals, 138 (2020), p. 109917, 10.1016/j.chaos.2020.109917.

[24] D. Ardila et al., "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," Nature medicine, vol. 25, no. 6, pp. 954–961, 2019.

[25] K. Suzuki, "Overview of deep learning in medical imaging," Radiological physics and technology, vol. 10, no. 3, pp. 257–273, 2017.

[26] N. Coudray et al., "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," Nature medicine, vol. 24, no. 10, pp. 1559–1567, 2018.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[28] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, "Regression models for prognostic prediction: advantages, problems, and suggested solutions." Cancer treatment reports, vol. 69, no. 10, pp. 1071–1077, 1985.

[29] P. Lapuerta, S. P. Azen, and L. LaBree, "Use of neural networks in predicting the risk of coronary artery disease," Computers and Biomedical Research, vol. 28, no. 1, pp. 38–52, 1995.

[30] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," American heart journal, vol. 121, no. 1, pp. 293–298, 1991.

[31] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.

[32] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," Plos one, vol. 15, no. 3, p. e0231236, 2020.