IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Analysis of COVID-19 Pandemic in India – Using Machine Learning Models

Shivakumar Nethani[1] and Dr. P. U. Anitha[2], P. Ramana[3]
Assistant Professor, Department of Computer Science and Engineering[1,2,3]
Hyderabad Institute of Technology and Management, Hyderabad, Telangana, India[1,3]
Christu Jyothy Institute of Technology and Sciencet, Janagaon, Telangana, India[2]
Correspondence should be addressed to Shivakumar Nethani; shivakumar.kumar093@gmail.com
Correspondence should be addressed to Anitha Podishetty; anithapodishetty123@gmail.com
Correspondence should be addressed to Rama Pochaboina; pochaboinarama@gmail.com

Abstract: *The early and reliable detection of COVID-19 infected patients is essential to prevent and limit its outbreak. The PCR tests for COVID-19 detection are not available in many countries, and also, there are genuine concerns about their reliability and performance. Motivated by these shortcomings, this article proposes a deep uncertainty-aware transfer learning framework for COVID-19 detection using medical Images. One of the practical and topical applications in the current scenario is to use the power of Machine Learning to study various aspects of the on going pandemic (COVID-19), since the entire world is in its grip. COVID-19 was declared a global pandemic on 11th March 2020 by WHO. Worldwide more than 43 Million people have contracted this viral disease and more than 1.1 million people have succumbed to it (as on 27th October 2020). Our approach uses Classification from Supervised Learning techniques to solve this problem. The efficacy of this approach could be used to scale and develop automated systems that could predict the likeliness of Covid-19 based on laboratory tests that are readily accessible. From the features presented to us in the dataset, we are able to predict with 87.0 - 97.4 percent accuracy at a 95 percent confidence level that a patient is suffering from Covid-19 when biomarkers are taken into consideration. The number of people affected by COVID-19 in India is increasing at a fast pace and currently India has the second highest number of cases and third highest casualties in the world. Four different Machine Learning algorithms namely Random Forest Regression, Multiple Linear Regression, Support Vector Regression and Lasso Regression have been considered. A Kaggle dataset consisting of figures of confirmed cases, patients recovered, and people that have died due to COVID-19 across India over a particular period of time has been used. The results of this study indicate that Random Forest Regression provides the most accurate results whereas Support Vector Regression is least accurate.*

Keywords: COVID-19, Lasso Regression, Multiple Linear Regression, Support Vector Machine, Machine Learning, Classification, Supervised Learning, Random forest Classifier

I. INTRODUCTION

A major ongoing crisis and the cause of around 1.1 million deaths worldwide, the pandemic COVID-19 is considered to be the greatest problem that has befallen the world since World War II. The first case of this disease is believed to have been recorded in December 2019 and its origins seem to have a link to a wholesale food market in Wuhan, the capital of Hubei Province in the People's Republic of China. The rapid spread of COVID-19 worldwide has claimed thousands of lives and has put unprecedented pressure on the healthcare systems around the world. The World Health Organization has emphasised the need for comprehensive testing in order to fight the virus. As of May 2020, it has impacted around 170 countries and regions. There are globally more than 4M identified COVID-19 cases and the number of death is fast approaching 300k. Lockdowns and restrictions have been applied by authorities in different countries to slow down its spread. The impact on the world economy has been massive due to restrictions applied to people's movement and the disruption of supply chains. Machine Learning, an application of Artificial Intelligence, has been used since decades in numerous fields and industries such as medical diagnosis, product recommendation systems,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

image processing, etc. Not many researches have been carried out to forecast the trend of the disease in India. The first patient of COVID-19 in India was observed on 30th January 2020 and the outbreak of the disease has created havoc throu- ghout the country. Many steps to contain the spread of the disease were initiated by the Government, including nation- wide lockdowns, educating the people on taking appropriate precautionary measures like the usage of face masks, disinfectants, etc. Although lockdowns are essential to curb the spread of the disease, the downside of this is that it has adversely affected the nation's economy, giving rise to unemployment, leading to the untimely deaths of numerous daily wage workers and migrant laborers. Machine Learning, an application of Artificial Intelligence, has been used since decades in numerous fields and industries such as medical diagnosis, product recommendation systems, image processing, etc. Not many researches have been carried out to forecast the trend of the disease in India. Therefore, in this paper an attempt has been made to forecast the pattern of death rate in India, over a period of 272 days, spanning a range of 9 months, using Supervised Machine Learning Regression models.

II. RELATED WORK

The need of the hour is to understand the spread of the disease and to be able to predict the number of cases that may arise, so that appropriate action can be planned. Machine Learning (ML) techniques have proved beneficial in making forecasts in different sectors like health, stock market and the weather. It has been proven to be of significant utility in the medical domain to detect and forecast the medical conditions of patients with respect to specific diseases like conditions of the heart, cancer, diabetes, etc. Attempts have been made by researchers to study COVID-19 through different techniques. The blood and clinical samples were collected of patients who visited the hospital for a suspected infection of Covid. The anonymized data was made available in a standardized and normalized form. It has a unit standard deviation and a mean of zero. This data hasn't captured a critical feature D-Dimer and has fewer values of potassium which have been seen to be of critical nature in relation to Covid. The studies showing the importance of these features were published at a later date and at the time, their respective importance was not completely known.
The features which comprise various clinical parameters can be broadly categorized under CBC, liver function, renal analysis, salt tests, blood gas analysis (arterial and venous), influenza tests.

III. METHODOLOGY

The COVID-19-India Dataset is analyzed to study the trend of the pandemic in India. A bar graph showing the state-wise number of confirmed cases in India as on 27th October 2020 has been plotted (Fig. 1). Only features considered to have significance have been taken into account for this analysis, and the remaining features have been removed, before applying pre-processing techniques to the dataset. The pre-processing technique performed here is "Standard Scaler" which is used to standardize the numeric attributes of the dataset. The data set is then divided into two subsets: Training Set (75%) and Test Set (25%). The problem at hand satisfies the criterion of a regression problem, as the output variable is a continuous or real value. Hence, the following four Supervised Machine Learning regression models have been applied:

3.1 Random Forest Regression

It is an ensemble ML algorithm that can perform both regression as well as classification [13]. It makes use of a technique known as, "Bootstrap Aggregation". Random Forest Regression is quite powerful and works better than other regression techniques in general. It is an ensemble of decision trees and the maximum depth is specified as well, to prevent the model from over fitting. The method of averaging it follows makes it more accurate than a decision tree. However, it faces one disadvantage - it is unable to form trends that would enable it to extrapolate values falling outside the training set.

Multiple Linear Regression

Multiple Linear Regression is a type of supervised ML algorithm which is widely utilized for predictive analysis . The inspection of Linear Regression depends on 2 values:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

- Independent Values
- Dependent Values

It is thus also used to find a linear relation between these values. It is recognized as "Multiple Linear Regression" due to multiple independent variables. The linear equation used to predict the output allocates one scale factor to each of the input values known as Coefficient and is represented by β1. Alongside this, an additional coefficient β0 known as intercept is also added to provide an additional degree of freedom. Mathematically, the equation is given as:

$$y=\beta 1x+\beta 0+\varepsilon \qquad (1)$$

Here, ε is the error term used to note the variation between both x and y. The goal is to find the optimal values of this coefficient and obtaining the regression line.

LASSO Regression

LASSO Regression also known as L1 Regularization, can be applied to a variety of models, and is mainly used for high-dimensional linear regression models, as the shrink age makes the model more stable and accurate . Extra features are penalized using regularization.

Problems having multi co-linearity of variables can be handled using LASSO Regression. The formula is given as:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{n=1}^{N} \frac{1}{2}(y_n - \beta x_n)^2 + \lambda \sum_{i=1}^{p} |\beta_i| \qquad (2)$$

The coefficient is set, which can be interpreted as min (sum of square residuals + λ |slope|), where λ |slope| is penalty term.

Support Vector Regression

It is a type of Support Vector Machine (SVM) algorithm that is not very popular, as SVM is mainly used for classification problems . SVR works so as to reduce the generalization error bound to produce a generalized performance. The objective of any regression model is to find a function which approximates mapping from an input domain to the real numbers on the basis of a training sample.

Thus, the main objective of SVR is to take into consideration the points that lie within the decision boundary line and to find the best fit line. The best fit line is the hyper plane that has a majority of the points.

Finally, the quality of the prediction models is evaluated using the following Evaluation Metrics:

- Mean Absolute Error
- Mean Square Error
- Root Mean Square Error
- R-Squared Score

A flow-chart of the steps followed in this problem has been presented in Fig. 1.

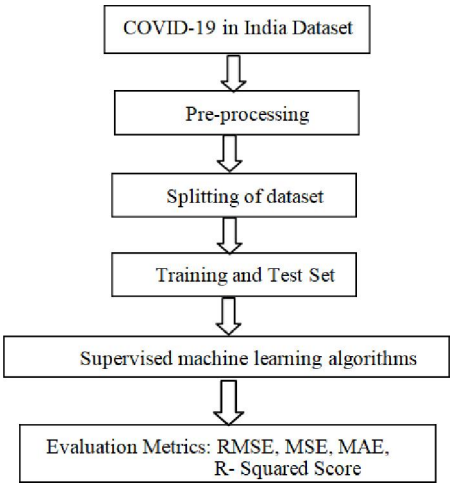

Fig. 1. Working Flow-chart

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

IV. GRAPHICAL ANALYSIS OF DATA

State-wise analysis of total confirmed cases

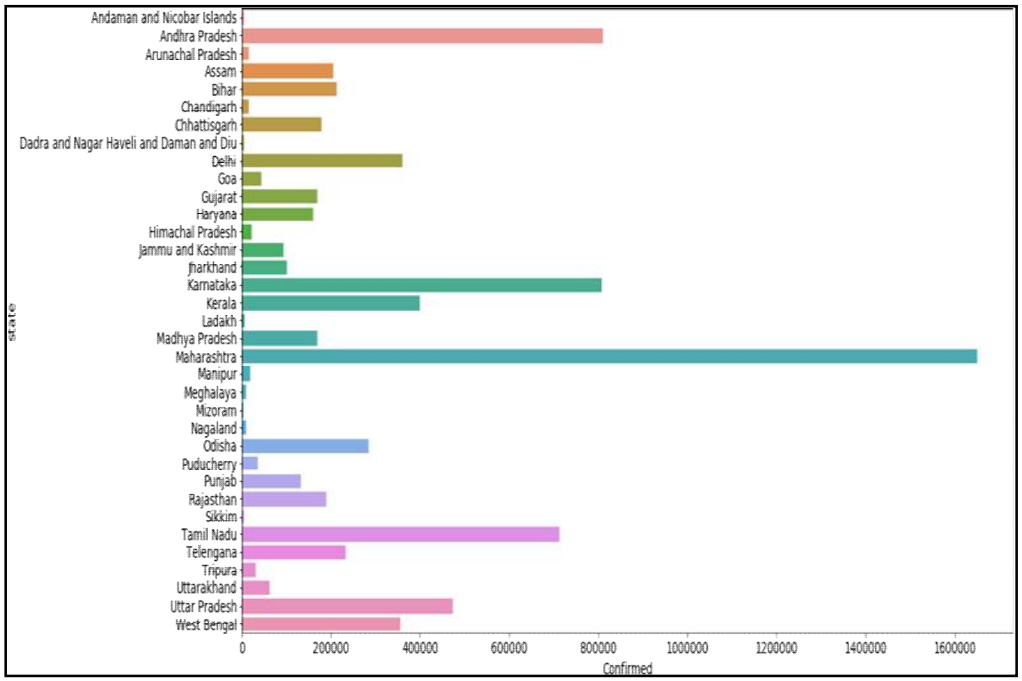

Fig. 1. State-wise number of confirmed cases in India as of 27th October 2020

Inferences drawn from Fig. 1 as of 27th October 2020 are listed below:

Maharashtra, by a huge margin has the highest total number of confirmed cases of COVID-19 in India.

Andhra Pradesh and Karnataka occupy the second and third positions, respectively.

Correlation Matrix Heat map of Features Used

Fig. 2. Correlation matrix heat map plotted between deaths, confirmed cases, patients cured and active cases

Inferences drawn from Fig. 2 are listed below:

The high values (close to 1) in all the boxes in the heat map denote the linear trend or strong positive correlation among the four features plotted which are:

- Number of Deaths
- Number of Cured Cases
- Number of Confirmed Cases
- Number of Active Cases

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

The positive correlation indicates that as one variable increases, so does the other and the closer the value is to 1, the stronger the relationship.

Pie chart of Features Used

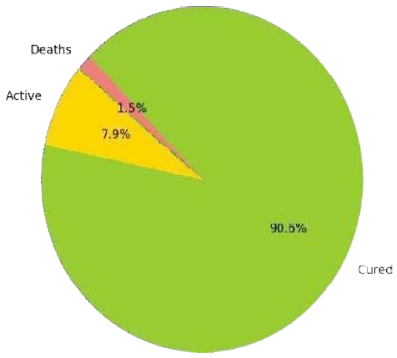

Fig.3.Pie chart denoting the distribution of deaths, active cases and patients cured as a percentage of the total confirmed cases

Inference drawn from Fig. 3 is listed below:

With a relatively low percentage of deaths (1.5 %) compared to the world average of 2.64%, India has been able to successfully contain the virus due to the various measures adopted.

V. CONCLUSION

Over the past nine months, COVID-19 has spread all over India at an extremely rapid pace, resulting in nearly 120 thousand deaths and around 7.9 Million confirmed cases. The pandemic had resulted in a nation-wide lockdown which led to deaths of innumerable people below the poverty line due to unemployment, exhaustion, and starvation. The race for a vaccine has intensified over the past few months, and it is clear that without a proper vaccine, this pandemic will only get worse. Using four Machine Learning regression models, the pattern of death rate over the past nine months has been analysed. The graph shows a clear indication of an extreme rise in the number of deaths. Visual representation of the state-wise increase in cases, pie chart and heat map has helped throw emphasis on the extremely dire situation. Upon evaluation of the four models using various evaluation metrics, it has been concluded that Random Forest Regression provides the best and nearly accurate results whereas Support Vector Regression shows the least accurate results. Multiple Linear Regression and LASSO Regression have also performed very well. The accuracy of the model could be increased with the addition of several attributes such as age, previously contracted diseases, etc. This work can be extended to improve the accuracy of prediction by doing further research as it is a problem with an ever-changing dataset and circumstances.

There are many rooms for improvement and further explo- ration. The performance of transfer learning algorithms could be majorly improved by fine-tuning them to extract more informative and discriminative features. Features obtained from different transfer learning models could be combined to develop hybrid models. Also, predictions from individual models could be combined to form ensembles. Last but not least, a state-of-the-art method could be applied for more comprehensive estimation of the uncertainty measures.

REFERENCES

[1]. (2020). Coronavirus Disease 2019 (COVID-19). Accessed: Feb. 27, 2020. [Online]. Available: https://www.who.int/emergencies/ diseases/novel-coronavirus-2019/situati%on-reports.

[2]. C. P. E. R. E. Novel, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China", Zhonghua LiuXing Bing Xue Za Zhi= Zhonghua Liuxingbingxue Zazhi, vol. 41, no. 2, p. 145, 2020.

[3]. A guide to WHO's guidance on Covid-19. (n.d.). Retrieved from https://www.who.int/news-room/feature-stories/detail/a-guide-to- who-s-guidance.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

[4]. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", IEEE Access, vol. 5, pp. 8869-8879,2017.

[5]. Emergency Approval to Use Diagnostic Test for Corona-Virus. Accessed: Jan. 23, 2020. [Online]. Available: https://www. reuters.com/article/us-china-health-cdc/u-s-cdc-seeks-emergency-%approval-to-use-diagnostic-test-for-coronavirus-idUSKBN1ZM2XS

[6]. Andre Filipe de Moraes Batista, J. L. (April-14-2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach.

[7]. A. Tomar and N. Gupta, "Prediction for the spread of COVID- 19 in India and effectiveness of preventive measures", Science of the Total Environment, vol. 728, 2020,138762.

[8]. B. Huang et al., "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," Lancet, vol. 395, pp. 497–506, May 2020.

[9]. Complete Blood Count.(n.d.). Retrieved from https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919

[10]. M. Barstugan., U. Ozkaya and S. Ozturk, "Coronavirus (COVID-19) classification using CT images by machine learning method", 2020 arXiv Preprint arXiv: 2003.09424.

[11]. S. Hassanpour et al., "Deep learning for classification of colorectal polyps on whole-slide images," J. Pathol. Informat., vol. 8, no. 1, p. 30, 2017.

[12]. Diagnosis of COVID-19 and its clinical spectrum. (n.d.). Retrieved from Kaggle: https://www.kaggle.com/einsteindata4u/covid19

[13]. Symptoms of Coronavirus.(n.d.). Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[14]. A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues, "Identifying pneumonia in chest X-rays: A deep learning approach," Measurement, vol. 145, pp. 511–518, Oct. 2019.

[15]. COVID-19inIndia, Available online: https://www.kaggle.com/sudalairajkumar/covid19-in- India? Select=covid_19_india.csv.