JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Applications of Image Processing and Machine Learning to Detect COVID-19

Chhaya Gupta, Kirti Sharma
Vivekananda School of Information Technology,
Vivekananda Institute of Professional Studies-Technical Campus
chhayagupta.spm@gmail.com, kirtisharmaa.11@gmail.com

Abstract

COVID – 19 (2019 novel coronavirus) which started in China had spread all over the world rapidly. It is the worst health crisis, the whole world has suffered since World War II. India reported its first COVID-19 positive case on 30 January 2020 in Kerala. It grew to 3 cases by 3 February 2020. The number of cases in India kept on increasing, and the virus is spreading widely. All necessary steps have been taken care of by the government of India to make sure that India is ready to face the challenge of the COVID-19 pandemic. The major step is to provide the right information to the citizens to take precautions. This paper is an effort to analyze the cumulative data of confirmed COVID-19 positives, deaths due to COVID-19, and recovered cases. This paper also examines the spread trend of this deadly virus in India. This paper compares different groups of people based on their gender and different age groups in India. This paper provides an overview of various Indian states with many COVID-19 positive cases, and all this data comes from the Kaggle online repository for datasets and the Ministry of Health & Family Welfare India. This paper uses Regression and classification machine learning models such as Linear Regression, Logistic Regression, Decision Tree Regressor, and Random Forest Classifier to analyse the data from the different datasets available to predict the area/state/union territory with an increase in the number of COVID-19 cases. The results show that the decision tree gave 99% accuracy in identifying COVID-19 positive cases.

Keywords: Coronavirus, Machine Learning, Decision tree Regressor, Logistic Regression, Linear Regression, Random Forest Classifier.

INTRODUCTION

COVID-19 pandemic first appeared in Wuhan, China in December 2019. It is a severe respiratory syndrome coronavirus that affected people worldwide [1][2]. This virus has been presumed to be transferred to humans from bats [3]. The virus tends to spread from one person to another. The infected person is treated in ICUs [4]. Many researchers have used chest CT scan images to identify the corona cases amongst normal person's chest CT scan images. Machine learning is getting used to identify the shreds of evidence of pneumonia using CT and xrays[5]. Bhattacharya et al. [6] in their paper have explored various DL techniques for processing CORONA medical photos. Moreover, they have also undertaken different use cases for covid detection using images. [7] A researcher has used CNN for rapid automatic COVID identification and shown good results. Expert systems [8]are made to categorize Xrays to recognize covid and pneumonia and are proven to show better accuracy when compared to other state-of-the-art models.

Symptoms of this disease are cough, fever, and respiratory disorders. It has been observed that men are more prone to this deadly virus than women as shown in fig 1 below. The figure is a visualisation of data of infected people based on their genders.

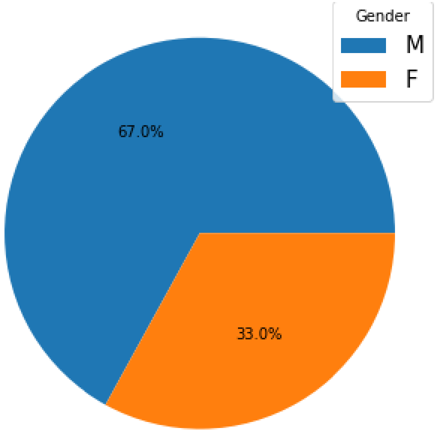


Fig. 1. Comparison of cases based on Gender.

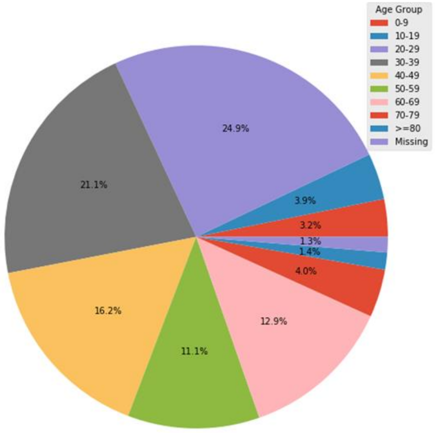


Fig. 2. Various Age groups percentage of cases.

In many developed countries, due to the unavailability of intensive care units, the health system has collapsed. The distribution of cases according to various age groups in India is shown in fig. 2. In the literature survey, there is no detailed study of the virus as of now. India is a big country with many states and there were very few testing kits available in India. But the Indian government made no stone unturned to make people safe whether it is the Lockdown period or strict rules for the implementation of social distancing. A 76-year-old man who returned from Saudi Arabia was the first victim in India, then the number of cases kept on increasing in the nation after 19 march 2020. The testing rates in India were among the lowest in the world and that was the main reason for an increased number of infections in India. The below graph in fig. 3 shows the number of confirmed cases, cured cases, and deaths in India.

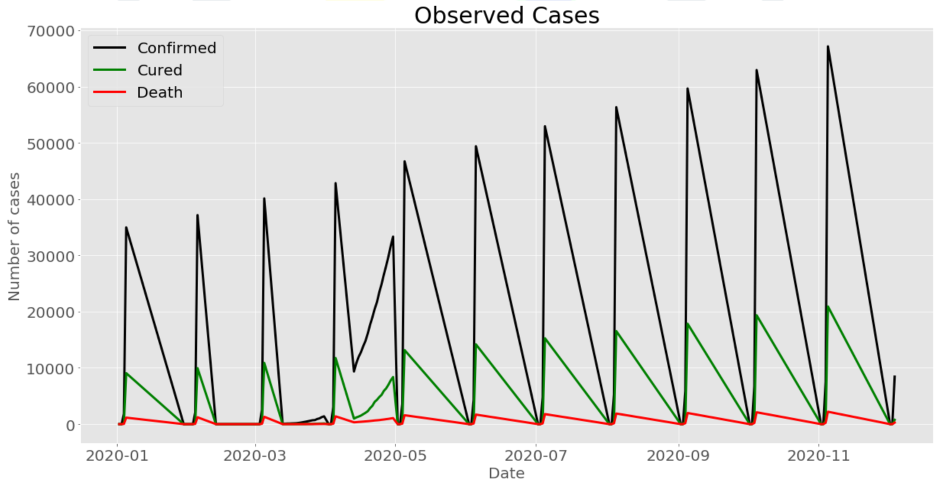


Fig. 3. Cases in India

After analysing the Indian states' dataset available, fig4 below shows the top 20 states with a number of confirmed cases, deaths, and cured cases respectively and table 1 gives the state-wise mortality rate in India. In this paper various machine learning methods have been applied to various Indian datasets available to estimate the accuracy of various machine learning methods available. The paper is divided further into Dataset Used, Machine Learning models used, and finally results and conclusions.

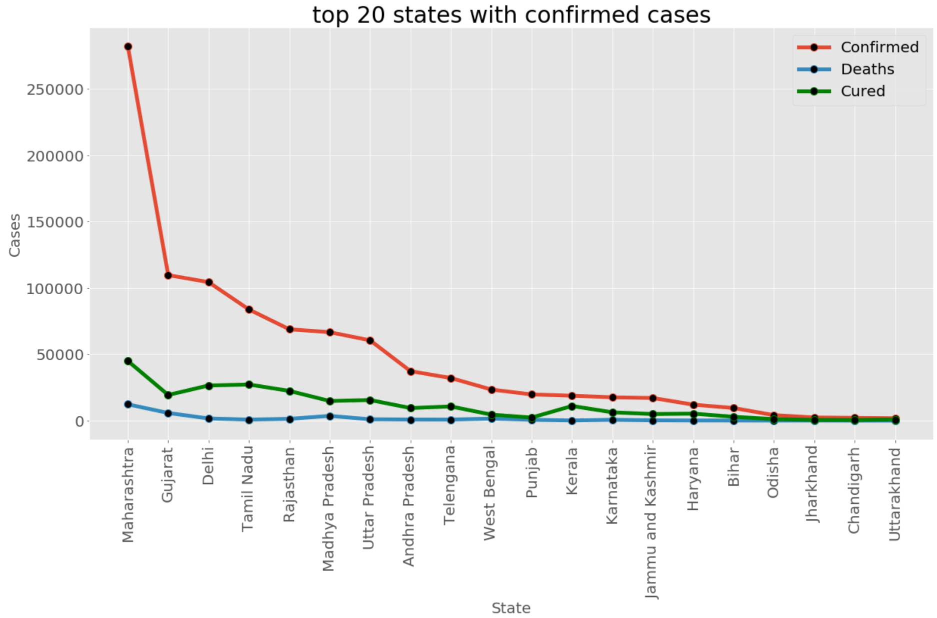
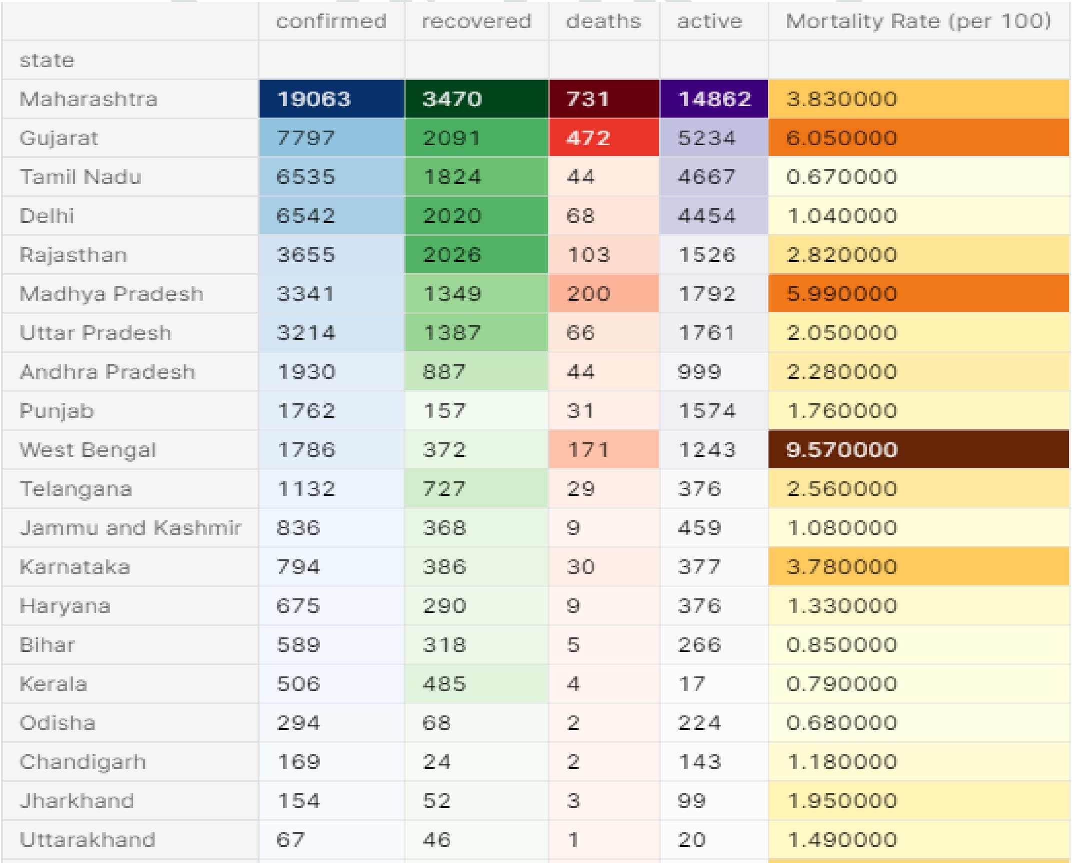


Fig. 4. Top 20 states Analysis

Table 1. Top 20 State-wise mortality rate.



	confirmed	recovered	deaths	active	Mortality Rate (per 100)
state					
Maharashtra	19063	3470	731	14862	3.830000
Gujarat	7797	2091	472	5234	6.050000
Tamil Nadu	6535	1824	44	4667	0.670000
Delhi	6542	2020	68	4454	1.040000
Rajasthan	3655	2026	103	1526	2.820000
Madhya Pradesh	3341	1349	200	1792	5.990000
Uttar Pradesh	3214	1387	66	1761	2.050000
Andhra Pradesh	1930	887	44	999	2.280000
Punjab	1762	157	31	1574	1.760000
West Bengal	1786	372	171	1243	9.570000
Telangana	1132	727	29	376	2.560000
Jammu and Kashmir	836	368	9	459	1.080000
Karnataka	794	386	30	377	3.780000
Haryana	675	290	9	376	1.330000
Bihar	589	318	5	266	0.850000
Kerala	506	485	4	17	0.790000
Odisha	294	68	2	224	0.680000
Chandigarh	169	24	2	143	1.180000
Jharkhand	154	52	3	99	1.950000
Uttarakhand	67	46	1	20	1.490000

Datasets Used

In this paper, different datasets have been used. The first dataset is COVID_19_India which describes various states and union territories of India with some confirmed, cured, and death cases respectively and with their attributes like age, gender, etc. Fig. 4 depicted in this paper is computed by analysing this particular dataset only. The next dataset used is Individuals_dataset which comprises details of individuals with their age, gender, current status whether still positive or not, and their state information. This dataset has been used to analyse the result that men are more prone to COVID-19 than women as shown in fig. 1. The next dataset used is Age_group details of individuals and it has been analysed to give the percentage chart for various age groups in figure 2. The testing dataset has also been analysed and compared with testing datasets of different countries like South Korea, Singapore, etc. The testing dataset below in Table 2 reveals that India was far behind in testing which was an important step to keep track of COVID-19 cases. All these datasets have been used from the Kaggle repository for dataset and the Ministry of Health & Family Welfare India. The dataset has been split into 75-25 ratios in training and testing sets respectively. Python 3.8 is the platform for the various analysis of data. All the datasets have been pre-processed to remove the missing values present in them. Matplotlib is the primary library used to analyse the data in the form of pie charts and tables.

Table 2. Comparison of Testing Done

Parameters	India	South Korea
Population	130 Cr	5.15 Cr
First Case reported on	28 January 2020	23 January 2020
Total Tests	1523213	6555194
Tests Per Million People	1171	9654
No. of Positives per 100 Tests (%)	4.56	2.44
Current Mortality Rate (%)	3.31	1.63

MACHINE LEARNING MODELS USED

This paper studies various machine learning models like Decision tree regressor, Linear regressor, Logistic regressor, and Random forest classifier. These models have been applied to the available datasets to understand which state will be predicting more COVID-19 positive cases in the future. All the models are compared based on their accuracy and the results clearly show that the decision tree is the best regressor to identify the data with 99% accuracy. This section also gives a brief description of all the methods used here. Firstly, this paper briefs about regression, classification, and then all the other methods that were used in this study.

- **Regression**: It is a supervised machine learning model that targets values based on independent variables. It is used to find the relationship between dependent and independent variables. It can be trained further to make predictions about real numbered outputs. The basis for regression is a hypothesis that can be either linear, quadratic, polynomial, non-linear, etc. Whenever there is a problem with real-time data, regression is the answer to that problem. Various regression models are Linear regression, polynomial regression, multiple Linear regression, etc.
- **Classification**: It is a supervised machine learning model which is used when we have categorical data. A categorical data can be "red" or "blue" or "yes" or "no" etc. It observes the given values and tries to draw some solution. Various classification techniques used are Decision Tree, Logistic regression, SVM, Random Forest, etc.

The following part of this paper briefs about the models used in this paper:

- **Decision Tree Regression**: It is a tree structure that breaks down a dataset into smaller and smaller subsets by developing a tree at the same time[9]. It is a built-in top-down format and the starting node is known as the root node. Standard deviation (SD) is used to calculate the homogeneity of numerical values available [10]. The standard deviation has the formula shown in equation 1,

  $$SD = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \qquad (1)$$

  Where x = given value

  $\bar{x} = \frac{\Sigma(x)}{n}$

  n = total number of samples

  The dataset is divided into further subsets based on standard deviation and the process repeats itself recursively.
- **Linear Regression**: It is an approach of supervised machine learning in which independent variables are focused on predicting the values. It is used for knowing the relationship between forecasting and

variables[11]. It finds a linear relationship between input and output, therefore the name is linear regression. The best fit line in this algorithm is represented in figure 5. The function for linear regression is explained in below equation 2 as,

Y = $\theta_1 + \theta_2 . X$ (2)

Where X = Input

Y = Output

θ_1 = Intercept

θ_2 = Coefficient of X

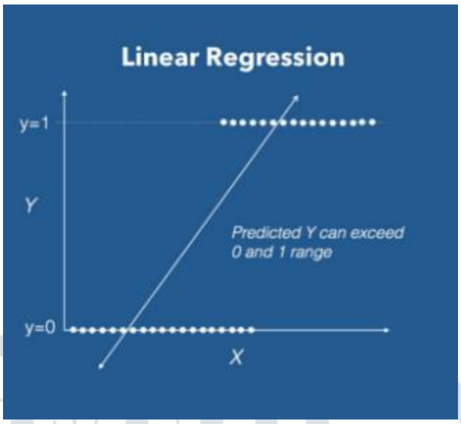


Fig. 5. Linear Regression

Once we have the best θ_1 θ_2 , then it provides the best fit line for our problem.

After finding the best fit line the model predicts the output Y and for this purpose cost function (J) is evaluated which is the Root Mean Squared Error (RMSE) between the predicted value and actual value [12]. Cost function (J) is calculated as in equation 3,

J = $\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$ (3)

And the motive is to minimise this error.

- **Logistic Regression**: It is a supervised learning classification that works with probability. It is used to predict the probability values of the dependent variable [13]. This model predicts P(X = 1) as a function of X. There are two types of logistic regression models:
  - Binary
  - Multi – Linear function

The cost function used by logistic regression is known as the Sigmoid function or logistic function, mathematically it is given by equation 4 below and the logistic regression curve is shown in figure 6. The cost function limit is between 0 and 1 [14].

F(x) = $\frac{1}{1+ e^{-(x)}}$ (4)

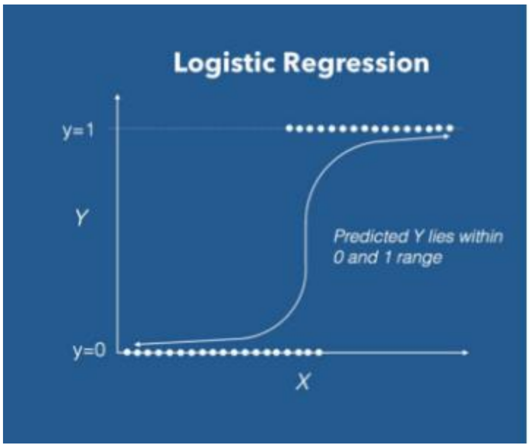


Fig. 6. Logistic Regression

The cost function in logistic regression is defined as:

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -log\big(h_\theta\ (x)\big) & if\ y = 1 \\ -log\big(1 - \ h_\theta(x)\big) & if\ y = 0 \end{cases} \tag{5}$$

- **Random Forest**: it is a supervised classification algorithm that is an ensemble learning model that uses a decision tree as its base model [15]. This model creates decision trees on the training dataset which forms a forest structure and prediction is done after it. This method is much more suitable for better results as it reduces the over-fitting problem [16]. Working of random forest model can be understood from the below steps:
Step 1: Random samples are selected from the dataset.
Step 2: For every sample, decision trees are created.
Step 3: Voting is done for predicted results.
Step 4: Most voted result is chosen as the final prediction result.
This model is useful for handling missing values in the given data and helps to avoid the over-fitting problem.

METHODOLOGY

This paper works on the datasets of COVID-19 provided by the Kaggle repository and the Ministry of Health & Family Welfare. Anaconda Spyder is the platform used with Python 3.8 version for implementing all the results. This study analysed different datasets available to make predictions by various machine learning models. Fig. 7 provides a flowchart of how the data has been pre-processed and how the results have been taken out. Step by step description of each step is provided below.

Step 1: Various datasets have been collected from mentioned websites.

Step 2: Each dataset has been analysed to give different results which are mentioned in previous sections of this paper.

Step 3: The dataset file having 1836 records contains 25 attributes like Gender, Age-group, State/Union territory, Cured, deaths, Confirmed cases etc. is used as the base dataset for making predictions.

Step 4: Data is splitted into training and testing sets in 75-25 ratio.

Step 5: Different Machine Learning Models mentioned in this paper are fitted to predict the results.

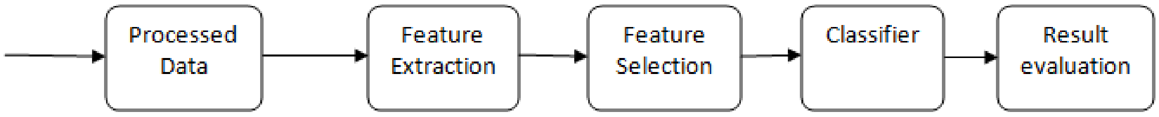


Fig. 7. Flow of Data

In this paper, a feature selection mechanism has been used to reduce the number of input variables. It is useful to reduce the number of input variables which reduces the computational cost of modeling and also improves the performance of the model. In this paper, the author has used SelectKBest method for feature selection. This is a sklearn feature selection method provided by python which takes different parameters and attributes like “score”, “pvalue” etc.

RESULTS AND CONCLUSION

Below fig. 8 provides the results of experiments performed on the dataset in a histogram format. The results clearly showed that the Decision Tree regressor provided 99% accuracy in predicting which state/union territory in India will have more cases and the answer provided was Delhi and Maharashtra with a tie. The decision tree and Random forest almost gave results near to each other.

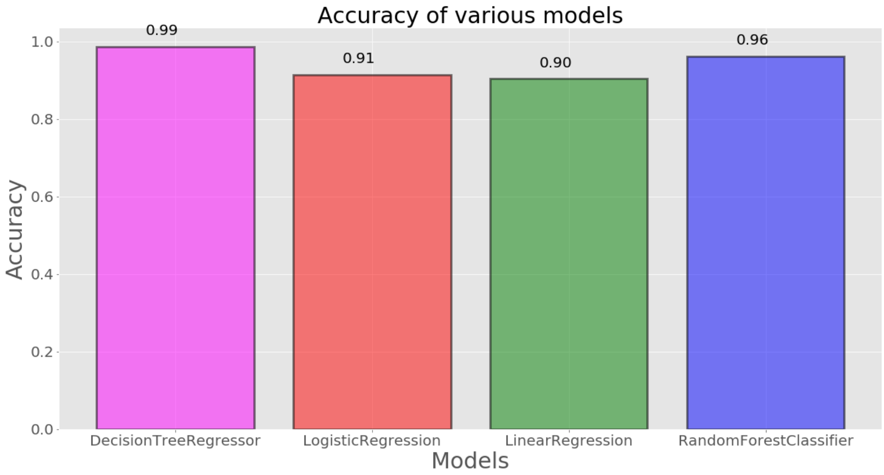


Fig. 8. Performance Evaluation of various machine learning models

This paper tried to identify which areas the Indian Government should look after for a long time. This study can help in dealing with uncertainty that may occur in the future and can thus prove a life savior. The entire world needs to understand the uncertainty of the future and should be ready with the resolutions of the same. COVID-19 is not yet over. They need to learn the correct information so better precautions can be taken. This paper doesn't look at all the aspects of varied datasets available for analysis but it is a step towards a better future. There is always room for improvement and here there is a big space.

REFERENCES:

[1] K. Roosa, "Real-time forecasts of the COVID-19 epidemic in China from February 5th to Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th , 2020," *Infect. Dis. Model.*, vol. 5, no. February, pp. 256–263, 2020, doi: 10.1016/j.idm.2020.02.002.

[2] A. Narin, C. Kaya, and Z. Pamuk, "Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks."

[3] C. Huang *et al.*, "Articles Clinical features of patients infected with 2019 novel coronavirus in Wuhan , China," pp. 497–506, 2020, doi: 10.1016/S0140-6736(20)30183-5.

[4] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," no. 5, pp. 1–10.

[5] F. Zhang, "Application of machine learning in CT images and X-rays of COVID-19 pneumonia," *Medicine (Baltimore).*, vol. 100, no. 36, p. e26855, 2021, doi: 10.1097/MD.0000000000026855.

[6] S. Bhattacharya *et al.*, "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey," *Sustain. Cities Soc.*, vol. 65, p. 102589, Feb. 2021, doi: 10.1016/j.scs.2020.102589.

[7] S. Liang *et al.*, "Fast automated detection of COVID-19 from medical images using convolutional neural networks," doi: 10.1038/s42003-020-01535-7.

[8] U. Subramaniam, M. M. Subashini, D. Almakhles, A. Karthick, and S. Manoharan, "An Expert System for COVID-19 Infection Tracking in Lungs Using Image Processing and Deep Learning Techniques," *Biomed Res. Int.*, vol. 2021, 2021, doi: 10.1155/2021/1896762.

[9] S. S. Rathore, "A Decision Tree Regression based Approach for the Number of Software Faults Prediction," vol. 41, no. 1, pp. 1–6, 2016, doi: 10.1145/2853073.2853083.

[10] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, "Decision tree regression for soft classification of remote sensing data," vol. 97, pp. 322–336, 2005, doi: 10.1016/j.rse.2005.05.008.

[11] G. Meyer, "Linear Regression under Fixed-Rank Constraints : A Riemannian Approach," 2011.

[12] D. Nguyen, N. A. Smith, and C. P. Ros, "Author Age Prediction from Text using Linear Regression," no. June, pp. 115–123, 2011.

[13] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, Y. Verbakel, and B. Van Calster, "REVIEW A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[14] A. I. Schein and L. H. Ungar, *Active learning for logistic regression : an evaluation*, no. April 2006. 2007.

[15] R. Joshi and M. Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm : Ensemble approach," pp. 426–435, 2017.

[16] P. Kaur, R. Kumar, and M. Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)," *Multimed. Tools Appl.*, vol. 78, no. 14, pp. 19905–19916, 2019, doi: 10.1007/s11042-019-7327-8.