Covid-19 Data Analysis and Visualization

Mamta Soni [2], Dr Sonali Ridhorkar [1]

[1] HOD, Department of Computer Science, G.H. Raisoni Institute of Engineering and Technology, RTMNU, Maharashtra, India 440033.

[2] Student, Department of Computer Science, G.H. Raisoni Institute of Engineering and Technology RTMNU, Maharashtra, India 440033.

Abstract— *Coronavirus disease 2019 is an infectious disease caused by serious acute respiratory syndrome coronavirus 2 ((SARS-CoV-2).. It was first identified in December 2019 in place Wuhan, China, and has resulted in an ongoing pandemic. The virus is primarily spread across people during close contact, most frequently via small droplets produced by coughing, sneezing, and talking. The droplets generally fall to the ground or onto surfaces rather than travelling through air over long distances. Less commonly, people may become infected by touching a dirty surface and then touching their face. It is most contagious in starting of first three days after the onset of symptoms, although spread is even possible before symptoms appear, and from people who do not show symptoms. The project will help in analyzing and recognizing the insights that will be gained by using the Technologies Python and Tableau are used to make all the visualizations which are displayed on the dashboard, these insights will help in identifying and giving an idea of how the number of covid cases are impacted as possibility of being diagnosed positive on the basis of the symptoms.*

Keywords- COVID-19, data analysis technique, Prediction, Machine Learning, Classification techniques.

I. INTRODUCTION

On 31st December 2019, in the city of Wuhan (CHINA), a group of cases of pneumonia of unknown cause was reported to World Health organization. In January 2020, a previously unknown new virus was identified, which is named 2019 novel corona virus. WHO has declared the COVID-19 as a pandemic. A pandemic is defined as disease spread over a wide range of geographical area and that has affected high proportion of the population.

Every person in the world suffers from the coronavirus, directly or indirectly. Someone is confronted directly, when the virus attacks them and some are indirectly affected because of the closure of their businesses, work, everyday work. Today, the global economy is also slowing down day in and day out. All countries are battling it, be it developing, developed or under development. Our goal is to make people aware so that they can protect themselves and unite the world to kill this disease and its existence.

As this COVID-19 is spread from person to person, Artificial intelligence based electronic devices can play a very pivotal part in preventing the spread of this virus. As the part of healthcare epidemiologists has expanded, the pervasiveness of electronic health data has expanded too. The increasing availability of electronic health data presents a major occasion in healthcare for both discoveries and practical applications to improve healthcare. This data can be used for training machine learning algorithms to improve its decision-making in terms of predicting the diseases. The project will help us in recognizing the insights that will be gained by using machine learning algorithms on the data, these insights will help us in identifying and giving an idea of how the number of covid cases are impacted as possibility of being diagnosed positive on the basis of the symptoms.

II. REVIEW

The different research spheres of data analysis that have considered COVID-19. While providing an epidemic computational model , GLEaM visualizes the spread of COVID-19 and analyzes realistic scenarios in comparison to data. As the model is developed, it regards transportation and interaction layers based on new emerging pandemic strains. Moreover, by allowing the integration of different processes not necessarily of biologic origin, the GLEaM model takes advantage of an individual's mobility to create simulations of the epidemic. Amidst large influenza-like illnesses, official health institutions may take weeks to reveal the data, hindering epidemiologic advancement. However, informal media typically holds available data in real-time which can allow for the development of epidemic forecasts.

Approximate Bayesian Computation (ABC) algorithms can be implemented to predict infectious disease trends when applied in a timely manner [6]. This paper provides training for users with little to no experience in parameter estimation from mathematical data. Three case studies with a focus on infectious diseases are presented to spotlight the many userbased factors that can increase accuracy and processing time.

Pujari, et al. proposed an approach that offers rapid and realistic epidemic predictions useful for health personnel in India. Although, the spread of COVID-19 is well-studied, most models are either computationally expensive, too coarsegrained to be reliable, or too fine-grained to be efficient. The model shown in this study is a hybrid approach of the Susceptible-Infected-Recovered (SIR) model. It includes well-mixed intra-city populations and intercity coupling based on transportation.

This allowed for rapid and accurate COVID-19 predictions in India. It predicted most of India's urban population to be exposed to the virus within the first 90 days of the epidemic unless strict preventative measures were taken that. The study concluded that a small infected population is sufficient for the rapid spread of the pandemic due to its incredibly infectious nature and the popular use of domestic transport networks.

III. PROBLEM DEFINITION

The pandemic has already taken grip over peoples' live. Since the pandemic has started, some countries are facing problem of ever-increasing cases. Through the data analysis of cases one can analyze how countries all over the world are doing in terms of controlling the pandemic. Analyzing data leads to accommodate the prevention model of the countries that are doing great in terms of lowering the graph. Predictions are made with the dataset which available to the individual/country/associations, therefore helping them to decide how far they are able to control the pandemic or up to how much extent they should guide preventive measures.

Through the project, a step towards helping people to understand the spread and predict the cases in their country has been done. The project also gives an vision of how a country is doing in terms of limiting the spread.

IV. METHODOLOGY

We are using Machine Learning to give predictions on the basis of data taken from government website[11], and then we clean the data by using excel cleaning methods and give prediction by using the algorithm with highest accuracy to predict COVID -ve or +ve on basis on 5 major symptoms.

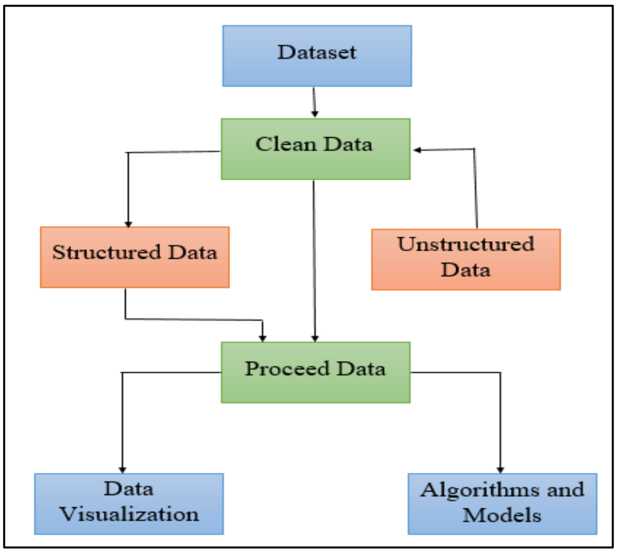


Fig [a]. Process of Data Collection

This is the total process of data collection, how the data was collected, and the method that is still using for increasing data collection. The whole process of data collection through graphs is going to be an idea.

The process can be explain in following given points :

1. First, Take the dataset, remove redundant data and organise the data according to our needs.

2. Second, Load the dataset on the Jupyter Notebook and apply data visualization techniques to understand the data better.

3. Third, then we calculate accuracy for various algorithms and plot a graph on the basis of accuracy of various algorithms.

4. Finally, using the accuracy graph we finally use the algorithm with best accuracy in this case (Decision Tree Classifier) to predict the person is either -ve or +ve on the basis of symptoms.

A. Description of the Process

We are building our own covid prediction

System using jupyter notebook.

We can describe the process in following steps:

Step 1: Cleaning the dataset

The very first step in our project is to get a reliable and authentic dataset for the prediction and analysis. Our search for dataset ended on which is govt website which has provided dataset for free use and is absolutely authentic.

Then next thing we did was to clean the dataset and remove unwanted columns from dataset for faster computation.

Step 2: Data Visualization

Here, we use the dataset and check the consistency of the dataset by checking the values out of the dataset randomly.

Then we do data visualization for better understanding of data by the use of various plots, graph and heat maps. All this graphs and plots gets us an insight into huge datasets easily.

Step 3: Computing Accuracy

In this step we compute accuracy of all the algorithms by checking the four algorithms mentioned here: Logistic Regression, KNN, Random Forest Classifier, Decision tree Algorithm , we selected these algorithms on the basis of their qualities of regression & classification.

Step 4: Predicting Covid +ve or -ve

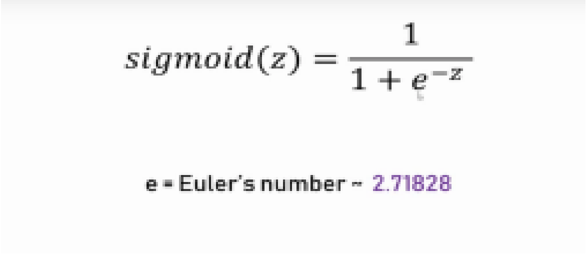
In the last step, all we need to do is plot a graph of accuracy of all the algorithms and use the algorithm with best accuracy to predict whether a person has corona or not.

We take input of 5 symptoms in binary values and using our predictor we predict the person is positive nor negative on the basis of these 5 symptoms.

B. Algorithm

1. Logistic Regression Logistic

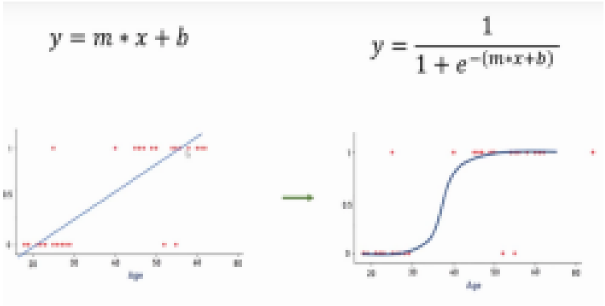
Logistic Regression is a Classification model, which tries to classify the data based on the probability of it occurring .This algorithm is used in multiple places where classification is required, we have used it to classify if the patient is susceptible to be infected by covid or not This is one of the classification methods which we have used. It used Sigmoid function to classify the data



$$sigmoid(z) = \frac{1}{1+e^{-z}}$$

e = Euler's number ~ 2.71828

Sigmoid function converts input into range 0 to 1



2. KNN

KNN is a supervised machine learning algorithm KNN forms groups based on the criterias and then decides for the incoming data where to put in which category It can be used for regression and for classification too, but mostly for the classification only its used.

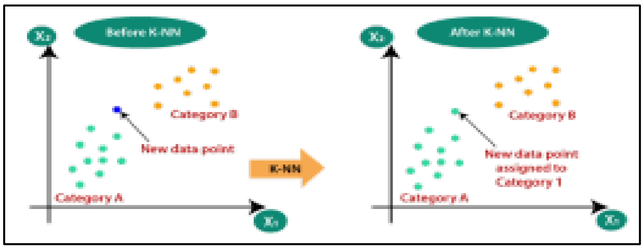


Fig [b]. Representation of KNN Algorithm

3. Random Forest Classifier

Random forest is a supervised learning algorithm.

The "forest" it builds is an group of decision trees, usually trained with the "bagging" system. The general idea of the bagging system is that a combination of learning models increases the overall result. Put simply: random forest builds multiple decision trees and combines them together to get a more accurate and stable prediction. One big advantage of random forest is that it can be used for both classification and regression problems, which form the most of current machine learning systems

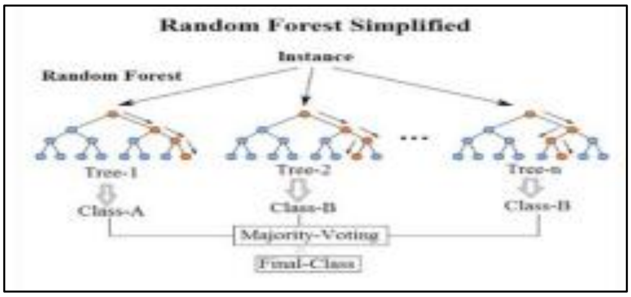


Fig [c]. Random Forest Classifier

4. Decision tree Algorithm

a. Decision Tree is a supervised machine learning algorithm

b.Two nodes which are decision node and leaf node are the ones making the decision

c.Repeated if clauses are at work when deciding the classification for the algorithm

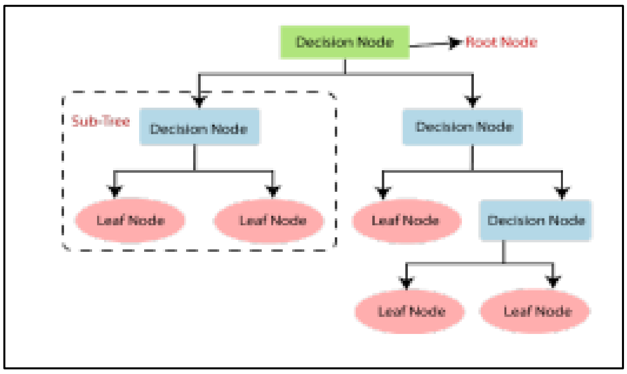


Fig [d]. Representation of Decision Tree

V. IMPLEMENTATION

1. *Covid-19 India's Data Analysis*

COVID-19 outbreak motivates to do an EDA on the datasets, scraped from different sources such as "Ministry of Health and family Welfare", "COVID-19 India website"and "Wikipedia" using "Python" and thus analyzing the spread and trend of the COVID-19 in India and done comparison with the neighboring and worst affected countries of the world. The dataset that uses EDA undergoes the method of normalization, choosing of essential columns using filtering, deriving new columns, and visualizing the data in the graphical format. This paper used "Python" for "data processing" and "web scrapping", "pandas" library to process and extract information from the available dataset. Appropriate graphs created for the better visualization are the results of "Matplotlib" and "Seaborn" library of Python.

1) COVID-19 Cases Spread in India over time

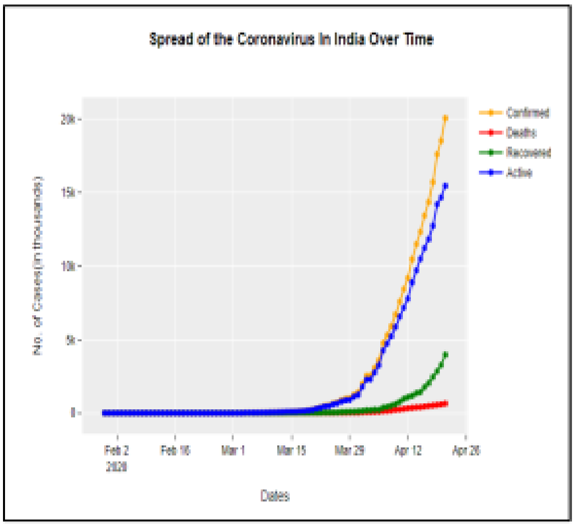


Fig [e]. Spread of COVID-19 cases in India over time

In the Figure 1, the X axis shows the Dates on an interval of 15 Days and Y axis shows the number of cases (in thousands).Orange line shows "Confirmed cases" (positive cases), Red line represents "Deaths" cases which showing the number of cases who had lost their lives, Green line represents "Recovered" cases which depicts the count of people who has recovered and the Blue line represents "Active" cases, the difference of Deaths and Recovered from Confirmed cases.

2) Vaccinated States in India

Using matplotlib, we create a histogram using the hist method. We used the histogram to show the details of the campaign vaccination that is currently running in India on a massive scale.

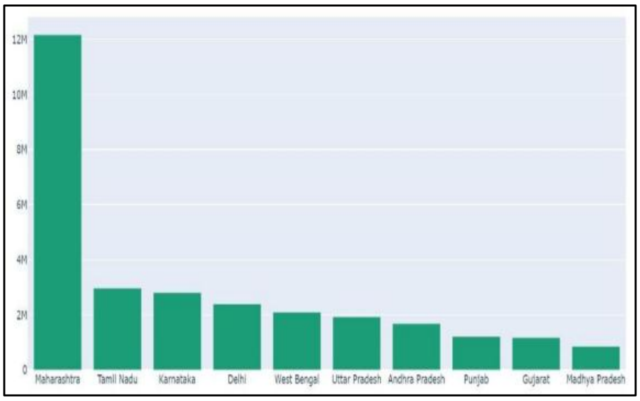


Fig [f]. India Vaccination Details

3) Age-wise spread of COVID-19 in India

The pie chart analyse the spread of COVID-19 in India to understand which age group is affected most.

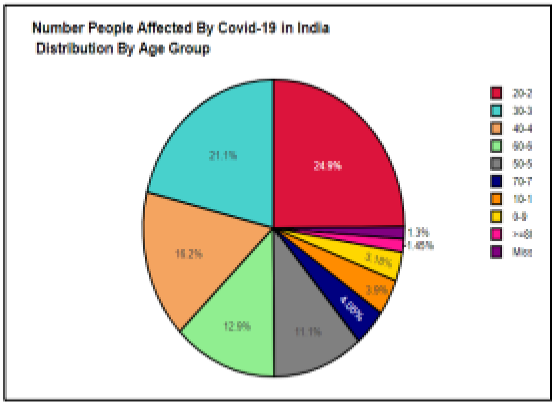
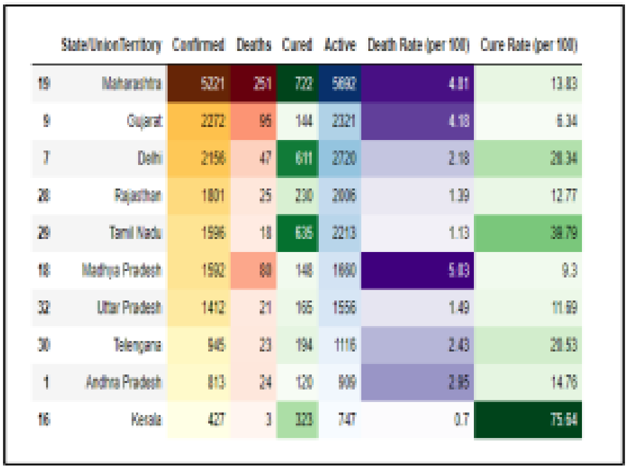


Fig [g]. Age-wise spread of COVID-19 in India

4) State-wise Analysis of COVID-19

Kerala reported the first coronavirus case in India on January 30 when a student who had returned from Wuhan. Till February 3, two more students were tested positive after their return from Wuhan. Till then the spread of the COVID19 in India has been on rampage



	State/UnionTerritory	Confirmed	Deaths	Cured	Active	Death Rate (per 100)	Cure Rate (per 100)
19	Maharashtra	5221	251	722	5692	4.81	13.83
9	Gujarat	2272	95	144	2321	4.18	6.34
7	Delhi	2156	47	611	2720	2.18	28.34
28	Rajasthan	1801	25	230	2006	1.39	12.77
29	Tamil Nadu	1596	18	635	2213	1.13	39.79
18	Madhya Pradesh	1592	80	148	1680	5.03	9.3
32	Uttar Pradesh	1412	21	165	1556	1.49	11.69
30	Telengana	945	23	194	1116	2.43	20.53
1	Andhra Pradesh	813	24	120	909	2.95	14.76
16	Kerala	427	3	323	747	0.7	75.64

Fig [h]. Sate-wise analysis of COVID-19 in India

5) Symptoms observed for COVID-19

The bar graph in Fig. 5 X-axis shows the percentages and Y-axis shows the name of symptoms which has been analyzed from the people who has been tested till now in India. This is the observation to keep the average symptoms checklist that has been created to keep a lead if some new patients comes and can be helpful in classifying them as positive and negative

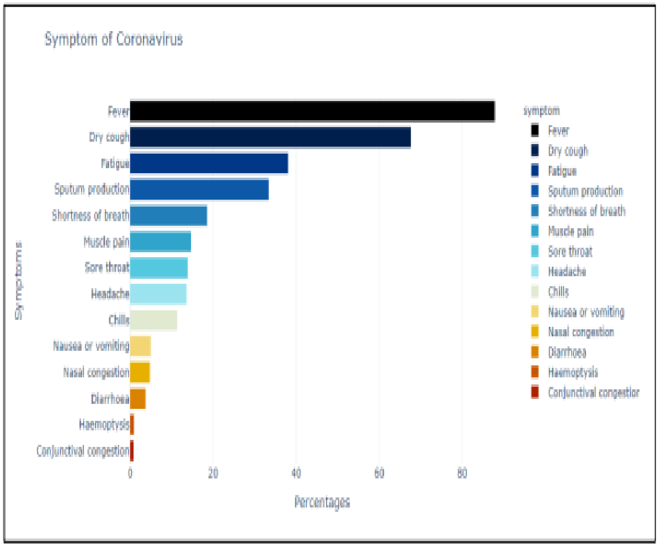
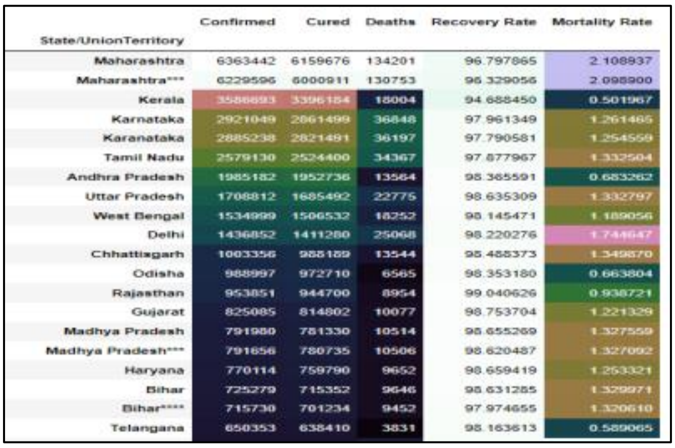


Fig [i]. .Symptoms for COVID-19

6) Recovery Rate

Recovery Rate is calculated by 100* (number of recoveries in a state/number of confirmed cases in a state)

The number obtain after this calculation is the number of recovered patients behind every confirmed cases in that state.



State/UnionTerritory	Confirmed	Cured	Deaths	Recovery Rate	Mortality Rate
Maharashtra	6363442	6159676	134201	96.797865	2.108937
Maharashtra***	6229596	6000911	130753	96.329056	2.098900
Kerala	3586693	3396184	18004	94.688450	0.501967
Karnataka	2921049	2861499	36848	97.961349	1.261465
Karanataka	2885238	2821491	36197	97.790581	1.254559
Tamil Nadu	2579130	2524400	34367	97.877967	1.332504
Andhra Pradesh	1985182	1952736	13564	98.365591	0.683262
Uttar Pradesh	1708812	1685492	22775	98.635309	1.332797
West Bengal	1534999	1506532	18252	98.145471	1.189056
Delhi	1436852	1411280	25068	98.220276	1.744647
Chhattisgarh	1003356	988189	13544	98.488373	1.349870
Odisha	988997	972710	6565	98.353180	0.663804
Rajasthan	953851	944700	8954	99.040626	0.938721
Gujarat	825085	814802	10077	98.753704	1.221329
Madhya Pradesh	791980	781330	10514	98.655269	1.327559
Madhya Pradesh***	791656	780735	10506	98.620487	1.327092
Haryana	770114	759790	9652	98.659419	1.253321
Bihar	725279	715352	9646	98.631285	1.329971
Bihar****	715730	701234	9452	97.974655	1.320610
Telangana	650353	638410	3831	98.163613	0.589065

Fig [j]. Recovery Rate

7) Male and Female Vaccination Ratio

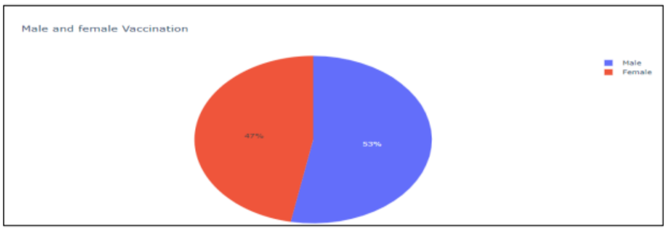


Fig [k]. Male and Female Vaccination Ratio

VI. CONCLUSION

In this study the main purpose was to analyze the COVID-19 spread in India since the day of outbreak and pattern of spreading of this virus. Study is done about the most common symptoms of COVID-19 that are observed till now, age wise spread of COVID -19 to observe which age group is affected most, the spread of the disease in India, the state wise trend of the pandemic to get detail understanding of how this is spreading. This analysis is to be fed into machine learning models for forecasting the number of confirmed cases, recovery cases and deaths across the globe by analyzing this COVID-19 dataset using machine learning algorithm. This project may be a better model in the future. Or the algorithm that is not giving good predictions, need to work on the algorithm so that the algorithm gives more good predictions. More models can try to create using algorithms. This research work, analysis, and prediction model will help this epidemic situation.

REFERENCES

[1] Zhu Junlan, Yang Chengke "Data-analysis-based discussion on COVID-19 Pandemic Shocks to the Economy and Policy Responses ," Management Science Informatization and Economic Innovation Development Conference, Year:2020

[2] Changchang hu, "The Topological Properties of COVID-19 Global Activity Time Series Forecasting," 5th International Conference on Information Science , Computer Technology and Transportation (2020)

[3] Ping Zeng, Kewei Yang,, " Using Big Data to Monitor the Impact of the COVID-19 Epidemic on Notifiable Diseases Reported in China," 6th International Conference on Big Data and Information Analytics, Year:2020

[4] Afshar Shamsi, Hamzeh Asgharnezhad, Shirin Shamsi Jokandan, "An Uncertainty-Aware Transfer Learning-Based Framework For Covid-19 Diagnosis," IEEE Transactions On Neural Networks And Learning Systems, Vol. 32, No. 4, April 2021

[5] Carson K. Leung, Yubo Chen," Big Data Visualization and Visual Analytics of COVID-19 Data," 24th International Conference Information Visualization, (2020)

[6] Huda Khaloofi, Jamil Hussain, Zahra Azhar,"Performance Evaluation of Machine Learning Approaches for COVID-19 Forecasting by Infectious Disease Modeling," International Conference on Women in Data Science at Taif University, Year:2021

[7] Yunxiang Liu, Yan Xiao," Analysis and Prediction of COVID-19 in Xinjiang based on Machine Learning," 5th International Conference on Information Science , Computer Technology and Transportation (2020)

[8] Abir Abdullha, Sheikh Abujar," COVID-19: Data Analysis and the situation Prediction Using Machine Learning Based on Bangladesh perspective," 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing , Year:2020

[9] Carson K. Leung, Yubo Chen, "Machine Learning and OLAP on Big COVID-19 Data," IEEE International Conference on Big Data, Year:2020

[10] Lyn Bartram, Michael Correll, Melanie Tory," Untidy Data: The Unreasonable Effectiveness of Tables," IEEE Transactions On Visualization And Computer Graphics, Vol. 28, No. 1, January 2022