International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Hybrid Machine-Learning Models for Predicting COVID-19

Dr. Harsh Mathur, Vikas Kumar

Associate Professor (CSE Department), Rabindra Nath Tagore University, Bhopal, Madhya Pradesh, India
Research Scholar(CSE Department), Rabindra Nath Tagore University, Bhopal, Madhya Pradesh, India

ARTICLEINFO

Article History:

Accepted: 13 March 2023
Published: 18 March 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
131-144

ABSTRACT

COVID-19 dataset comprises time, nation, established cases, no. of recovered people, overall mortality rate. The data is integrated with climate data consisting of dampness, dew, ozone, awareness, highest and lowest temperature etc. Several online websites are present to collect the data. Some of these websites include “World meters”, “Our World in Data”, “World Bank Open Data”, and the official website of the World Health Organization (WHO). Moreover, researchers focus on human development reports for collecting other kind of data. The artificial intelligence based COVID-19 diagnosis strategies can generate more accurate results, save radiologist time, and make the diagnosis process cheaper and faster than the usual laboratory techniques. The COVID-19 detection has many stages like pre-processing, feature extraction, classification and performance analysis. In this work, a voting classification method is designed for the covid-19 prediction. It is analysed that proposed model increase accuracy, precision and recall for the covid-19 prediction.

Keywords : Naive Bayes, Voting Classifier, Logistic Regression, Precision , Recall

I. INTRODUCTION

Various diseases are occurred due to diverse kinds of pathogens which focuses on transmitting these disorders among individuals and animals. They can spread by several means with fast transmission speed. It is essential to detect, prevent and control the communicable infections at premature stage. At first, a secretive viral pneumonia was found in Wuhan, China. WHO states this virus as 2019-nCoV that leads to cause COVID-19. At present, this virus is become 7th species of coronavirus which has potential to infect human beings. COVID-19 and SARS diseases are occurred due to the coronaviruses which were firstly found in 2003. The symptoms of these disorders are almost same but have little bit difference . Individuals having the 2019-nCoV infection generally experience varying symptoms, such as fever, mild cough, or pneumonia, that sometimes lead to death. The death rate of corona virus disease 2019 is about 2% to 4%. Though, this percentage is inaccurate and a variation can be seen as more information is present.

Meanwhile, it is impossible to mitigate the severity of this virus. The individuals having this virus meet diverse consequences. A research programme is organized at huge scale by various countries in order to prevent and control this disease due to its spread at global level.

COVID-19 Risk Prediction Framework

This framework has 6 modules which are executed for different purposes such as to collect the data, pre-process the data, select the features, develop the risk predictive framework, validate this model, and nomogram development and probability of COVID-19.

Figure 2represents the general Covid-19 risk Prediction Framework.

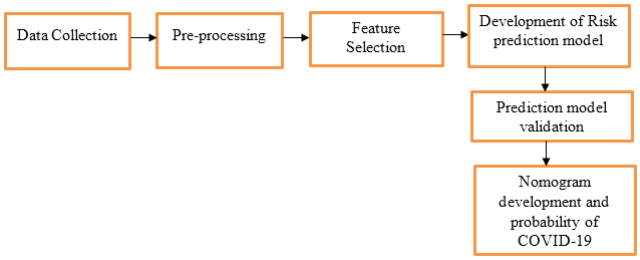


Figure 1: COVID-19 risk prediction framework

Diverse phases of this model for predicting corona disease are defined as:

a. Data collection: COVID-19 dataset comprises date, country, confirmed cases, recovered cases etc. The data is integrated with climate data consisting of humidity, dew, ozone, perception, max temperature, etc. . Numerous websites are present to collect the data. Some of these websites include “Worldometers”, “Our World in Data”, “World Bank Open Data”, and the official website of the World Health Organization (WHO).

b. Data Pre-processing: A huge volume of noise, misplaced, and variable data is contained in the real time datasets. The major element for analysing the process of DM is the quality of data in order to predict and detect the disease effectively. The data of lower level leads to provide poor and imprecise predictive results. Therefore, the several phases are executed to offer more efficiency and aptness to the dataset for predicting the data.

c. Feature Selection: The hospital admission and ventilator treatment is predicted with regard to age and BMI factors for all targets and these factors play an imperative role in predicting the disease. Hypertension is another factor to predict ICU admission, etc. Age, dementia, are major factors for the occurrence of disease in the hospitalized patients.

Literature Review :

Ming-Yen Ng, et.al (2020) presented two validated risk prediction algorithms for COVID-19 positivity for which readily available parameters were considered in a general hospital setting [1]. The clinical utilization was facilitated using nomograms and probabilities. The patients having COVID-19 or normal were taken from the four hospitals of Hong Kong. The algorithms were generated with the help of MLR (Multivariable logistic regression) and its validation was done in H-L (Hosmer-Lemeshow) and calibration plot. The evaluation of nomograms and probabilities was performed for quantifying the different parameters such as sensitivity, specificity, PPV and NPV. It was analyzed that a superior sensitivity and NPV were found at lower probabilities and superior specificity and PPV were obtained when the probabilities were high.

Alberto Utrero-Rico, et.al (2021) designed a mortality prediction framework in order to predict the patients who were hospitalized due to COVID-19 [2]. This framework was utilized for computing the probability

of death with regard to lactate dehydrogenase, IL-6, and age. Three validation cohorts were put forward to quantify the discrimination and calibration. The individual risk factor effects were re-estimated in the overall cohort to update the designed framework. In the first two cohorts, this framework performed efficiently and the third cohort represented the excellent calibration. The updated framework was also assisted in predicting the fatal outcome in patients without respiratory distress at the time of evaluation.

Arjun S Yadaw, et.al (2020) intended an accurate predictive model of COVID-19 mortality in which unbiased computational techniques were utilized and the clinical attributes were also recognized [3]. The analysis related to development and validation of predictive model included the implementation of ML techniques for clinical data which was taken from a huge cohort of patients suffered from corona virus disease 2019 and treated at New York City for predicting the mortality. The mortality was predicted on the dataset using the intended model which was planned on the basis of clinical attributes and patient characteristics. The intended model provided the accuracy around 0·91.

Victor M. Castro, et.al (2021) suggested the supervised ML to EHR data taken from 3 hospitals where the patients suffered from coronavirus disease 2019 were admitted [4]. Using this data, an incident delirium predictive framework was constructed. Those hospitals were considered for authenticating the framework. The c-index was found 0.75, when the suggested framework was implemented in the external validation in which 755 patients were comprised. It was observed that the suggested framework provided the sensitivity around 80%, its specificity was computed 56% and negative predictive value was found 92%. This approach performed similarly in case of subsamples including age, sex, race for critical care and care at community as well as academic hospitals.

Ahmad Sedaghat, et.al (2020) presented an SEIR-PAD technique for determining the populations that are diseased and susceptible to infection [5]. This algorithm involved seven sets of ordinary differential equations with eight unknown coefficients. The MATLAB was utilized for solving these coefficients in numerical manner. For this purpose, an optimization algorithm was executed for employing four-set data of COVID-19 in which cumulative populations of infected, deceased, recovered, and susceptible were comprised. The outcomes demonstrated that the introduced algorithm offered insight to deal with COVID-19 pandemic in GCC countries.

Anwar Jarndal, et.al (2020) designed a model to anticipate the number of deaths from corona virus disease in 2019 based on the number of cases of elderly, diabetic, and smokers that have been documented [6]. Using the GPR (Gaussian Process Regression) method, this model was created. A comparative analysis was conducted on the developed model and ANN (Artificial Neural Network). A reliable data, that the World Health Organization (WHO) had published, was utilized to implement this model. The outcomes depicted that the developed model was adaptable to predict the number of deaths occurred because of coronavirus disease 2019. Furthermore, this model was also assisted in preparing effective measures so that the number of deaths was reduced.

D. Haritha, et.al (2020) established a TL (transfer learning) paradigm for coronavirus prognosis using chest X-rays [7]. Using CNN's algorithm GoogleNet, the image was categorised. This model was capable of classifying the images positively which determined whether the COVID-19 was present. The results indicated that the accuracy attained in training using the recommended model was calculated 99% and

accuracy in testing phase was computed 98.5% while predicted the corona disease. In the remote places that had not any experienced practitioners, the primary health workers made the utilization of the recommended model.

Yifan Yang, et.al (2020) employed Long Short-Term Memory (LSTM) technique in order to anticipate the infected population in China [8].Moreover, this approach was unable to accurately capture the dynamics of the diffusion process, and it was shown that long-term predictions had a greater error rate. Susceptible-Exposed-Infected-Recovered (SEIR) was later proposed as a way to track the COVID-19 dissemination pathway. To accurately estimate the parameter and forecast the infected populations, a sliding window technique was helpful. The projected approach was useful for the epidemiological studies to understand the spread of the virus.

Manoj Kumar, et.al (2020) examined the hybrid epidemic SIR (susceptible-infected recovered) model for corona virus prediction [9]. The analysis of patterns and trends in the data led to the discovery of a logistic infection rate after taking into account various parameters when building the proposed model. The availability of ICU beds and ventilators per 100,000 people was then taken into account. The schedule to shift the phase of the corona virus had been offered by a variety of trustworthy sources. When predicting the corona illness, the examined model was able to increase accuracy with an R2 value of about 96.8%.

OnderTutsoy, et.al (2020) SpID (Suspicious-Infected-Death) system that was elevated, parametric, and had more dimensions [10]. Nonetheless following the constraints were lifted, analysis of the data from coronavirus showed that its dynamics were unstable. If nothing was done, harm would have resulted from this. Under the current scenario, thousands of days were needed to reduce the number of suspicious persons. Infected persons and fatalities would cease to exist after roughly 300 days. The developed technique helped identify those who contracted coronavirus in Turkey.Additionally, this system was trained using data from other nations, and it was able to forecast the equivalent coronavirus fatalities.

Nanning Zheng, et.al (2020) devised an ISI (improved susceptible-infected) system to predict different infection rates in order to examine the laws of transmission and development trends [11]. Thereafter, in an endeavour to include the LSTM and NLP section into the pre-existing paradigm, the implications of control and treatment along with the public's growing knowledge of avoidance were thought about. A hybrid AI (artificial intelligence) platform was created using this to forecast the coronavirus illness in 2019. The experimental findings showed that, in contrast to traditional models, the created hybrid model was flexible for reducing the error rates of predicting outcomes.

Yuling Zou, et.al (2020) concentrated on using the enhanced SEIR (Susceptible-Exposed-Infected-Recovered) to predict the coronavirus transmission trend at the early stages [12]. By adding more detailed conditions based on the conventional system, the actual situation of disease propagation was more accurately recreated. Real-time data and associated parameters were used to accomplish this. The results showed that, in comparison to the traditional model, the proposed system has the potential to forecast the real propagation trend with more accuracy. The more advanced model offered greater advantages to the advancement of the mathematical model area and better recommendations for governmental organisations to control the coronavirus.

ErtuğrulKaraçuha, et.al (2020) presented a dataset for which the previously recommended method was used [13]. Following that, two methods—Deep Assessment Mechanism and LSTM—were developed to forecast

the COVID-19. For anticipating the pandemics near future, a Gaussian prediction method was proposed, and its predictive ability was evaluated. The correlation coefficients and wavelet-based denoising were ultimately employed to assess the influence of history. The Deep Assessment Methodology's findings revealed that the average errors for confirmed cases, recovered cases, and death cases were up to 0.6671 percent, 0.6957 percent, and 0.5756 percent, respectively.

Yixiao Ma, et.al (2020) created an enlarged system known as eSEIR, where the well-mixed SEIR framework was upgraded on infection trends to estimate the coronavirus [14]. Incorporating an optimization technique led to the determination of the parameters. The suggested system was verified using epidemic data from China and Italy in order to reduce the RMSE ofthe expected curves. The findings revealed that the proposed approach was useful in predicting the disease's future spread in the US. Using the suggested system, it was possible to anticipate how the pandemic would expand in the future in the United States.

Saud Shaikh, et.al (2021) employed linear and polynomial regression models [15]. To quantify these models, the R squared score and error values were taken into account. Various confirmed, recovered, and death cases were predicted using the COVID-19 dataset for India based on the data that was available. In order to forecast the future trend of these situations, tableau's time series technique was used. This approach also offered future predictions for the total number of confirmed cases.

S. Tabik, et.al (2020) created the COVIDGR-1.0 database, which encompassed all coronavirus sensitivity levels [16]. This database was balanced and homogeneous. There are 426 positive and 426 negative PA-CXR pictures in this sample. To improve the generalisation ability of classifiers, a system known as COVID-SDNet (Smart Data based Network) was established. The intended course of action had succeeded, and its accuracy was evaluated to be 97.72% 0.95%. The coronavirus may have been predicted using this method.

Md Masud Rana, et.al (2020) studied a novel Internet of Things (IoT)-based secure communication system [17]. By creating and implementing an ideal SP (signal processing) method, COVID-19 was anticipated. On the basis of the specified gain, the dynamic system forecasting error was reduced in order to construct a reliable COVID-19 prediction method. The results confirmed that the researched strategy was effective for quickly forecasting this disease's states. The built simulator and analysis therefore proved to be a useful tool for anticipating the COVID-19 state, pre-emptive action, as well as for information security and privacy.

Narayana Darapaneni, et.al (2020) used a mathematical model that was predicted to track the SEIRD model, which included five components—Susceptible, Exposed, Infected, Recovered, and Deaths—was used to construct the epidemic curve for India [18]. The ARIMA and LR frameworks were employed to determine the R2 value for the Indian dataset. This dataset contained information from 7,553,182 patients diagnosed and 423,349 recorded mortality. The analysis revealed that the anticipated paradigm accurately forecasted how many verified COVID-19 cases would occur in the future days.

Ruizhi Han, et.al (2020) indicated using a BLS (Broad Learning System) to forecast COVID-19 patients' death using blood samples [19]. To quantify three systems, 375 patient blood samples were used. The suggested system had improved AUC, accuracy, and precision along with 94.80% specificity and 94.50% sensitivity. The results demonstrated how the suggested system outperformed competing systems. Future research will focus on doing analysis on a

sizable collection of infected blood samples in order to increase the accuracy of coronavirus prediction.

Sina Ardabili, et.al (2020) created a method known as ANN-GWO with the goal of predicting COVID-19 outbreak in which ANN (artificial neural network) and GWO (grey wolf optimizer) were combined [20]. The Global dataset was used to do this. The procedures and approach were trained, tested, and verified using the time-series data in relation to the MAPE (mean absolute percentage error) value. The findings showed that the tried-and-true method had provided 6.23% MAPE for training, 13.15% for testing, and 11.4% for authentication phases. Additionally, this method handled the task of prediction.

Naresh Kumar, et.al (2020) for the purpose of forecasting the temporal data of corona virus transmission, the ARIMA and Prophet time-series architecture were created [21]. Using the proposed system, the COVID-19 outbreak trends were identified as well as epidemiological stage data relevant to accepted nations. According to the investigations, the autoregressive integrated moving average model was more effective at predicting the prevalence of COVID-19. The governments found the forecast results useful when formulating plans to stop the virus's spread.

Xiaoyi Fu, et.al (2020) developed a method for coronavirus transmission forecasting using the recognition and surveillance of notable social media incidents [22]. This system includes a pipeline for creating COVID-19-related Event-Centric Knowledge Graphs from Twitter data streams. On the basis of the simulation of epidemic dynamic models utilising graph statistics, this disease was predicted more accurately. The trials for 128 nations were conducted using the Corona dataset from Johns Hopkins University. The results showed how effective the suggested system was. Additionally, this system had provided instructions for organising a company's return to work.

Yasin Khan, et.al (2020) used a CNN system to distinguish between COVID-19 patients, persons with respiratory infections, and healthy individuals in chest X-ray pictures [23]. To improve the data and foresee the COVID-19 disorder, the Monte-Carlo tool was employed to simulate the approach on the primary data patterns having confirmed cases and mortality rates. Moreover, LR of Gaussian mixture model was presented forforecasting the corona disease. The classification accuracy models were trained and tested using datasets from Kaggle and the University of Montreal. The findings showed that the recently implemented algorithm produced accuracy of about 100% and 96.66% during training and testing.

Andi Sulasikin, et.al (2020) a statement made that for estimating the quantity of coronavirus cases in Jakarta, Holt's exponential smoothing and ARIMA time-series models were built [24]. To determine the best models to forecast confirmed instances, several models were compared. The constructed models produced promising results and accurate forecasts that aided data-driven policy in public health and epidemiology. When forecasting the disease, the MSE and RMSE values were found to be lower and the R-Squared value obtained from the ARIMA method to be greater.

Chamara Sandeepa, et.al (2020) explained how the BLE (Bluetooth Low Energy) and GPS were used to build a social interaction tracking system [25]. Based on the data gathered, an algorithm was created to predict the likelihood of COVID-19 infection. Ultimately, a version of the system was delivered using an app and a web device. Additionally, a simulation for the investigation of the behaviour of the established algorithm was carried out using a graph-based model. The simulation showed that self-isolation was crucial to halting the progression of this illness.

II. Proposed Work

In December 2019, Wuhan, a province of China, was the coronavirus's (COVID-19) original location. More than 95 million cases had been reported globally as of January 2021, with a fatality rate of 2% of all closed cases. The pandemic's rapid spread is a global problem that poses a serious threat to both human health and the global economy. The 2019 coronavirus disease (COVID-19) is caused by a coronavirus. It has been dubbed a pandemic by the World Health Organization. Corona virus forecasting models use epidemiological data that includes the total population and the number of people who have already been infected.The latency duration and healing probability are two different parameters that are used to forecast infection patterns. However, these models are ineffective in capturing the various socio-economic and societal elements that have an impact on the virus's trajectory. The techniques need to propose which can predict coronavirus

III. Research Gap

The following is a discussion of the gaps in this work:

1. The dataset used to forecast the COVID-19 epidemic is intricate in design. As a result, correlations between various traits are difficult to establish. When there is no appropriate relationship between features, it can be difficult to identify the relevant features for COVID-19 prediction.

2. No outlier elimination technique was used in previous study to increase prediction accuracy. The dataset will be cleaned and classified for advanced processing after outlier removal.

3. Standard techniques for cleaning the dataset, such as under- or over-sampling, should be presented during the pre-processing step. The effectiveness of the prediction analysis process is increased by thorough dataset cleaning.

IV. Problem Formulation

The key focus of this work is to use data mining techniques for predicting covid-19. There are mainly three steps involved in the prediction. The first step of pre-processing is applied for removing missing, unnecessary values from the existing dataset. The next step establishes a relationship between feature and target set. The overall data is separated into two sets of training and testing in the final step. This work performs the task of covid-19 forecasting by applying three classifiers including RF (Random Forest), C4.5, and MLP (Multilayer perceptron). The outcome generated by these classification models is applied as input to the ensemble classifier for predicting covid-19 diseases. This work considers three performance metrics for analysing the efficiency of ensemble classifier. The obtain outcomes indicates about the intricacy of this classifier which should be reduced for making the forecasting of covid-19 possible.

V. Objectives

The key objectives of this work are discussed here:

1. To review and examine different types of data mining-based covid-19 prediction algorithms.

2. To apply an ensemble classifier for making the prediction of covid-19 possible.

3. To construct a data mining-oriented hybrid classifier for predicting covid-19 disorders.

4. To apply new approach and makes its compare with the former approach of covid-19 prediction w.r.t to different performance parameters.

VI. Research Methodology

There are many risk factors that may lead to covid-19.

Following are the various phases of covid-19 prediction:

A. Data Acquisition: The data is collected from various clinical organizations to perform experiments.

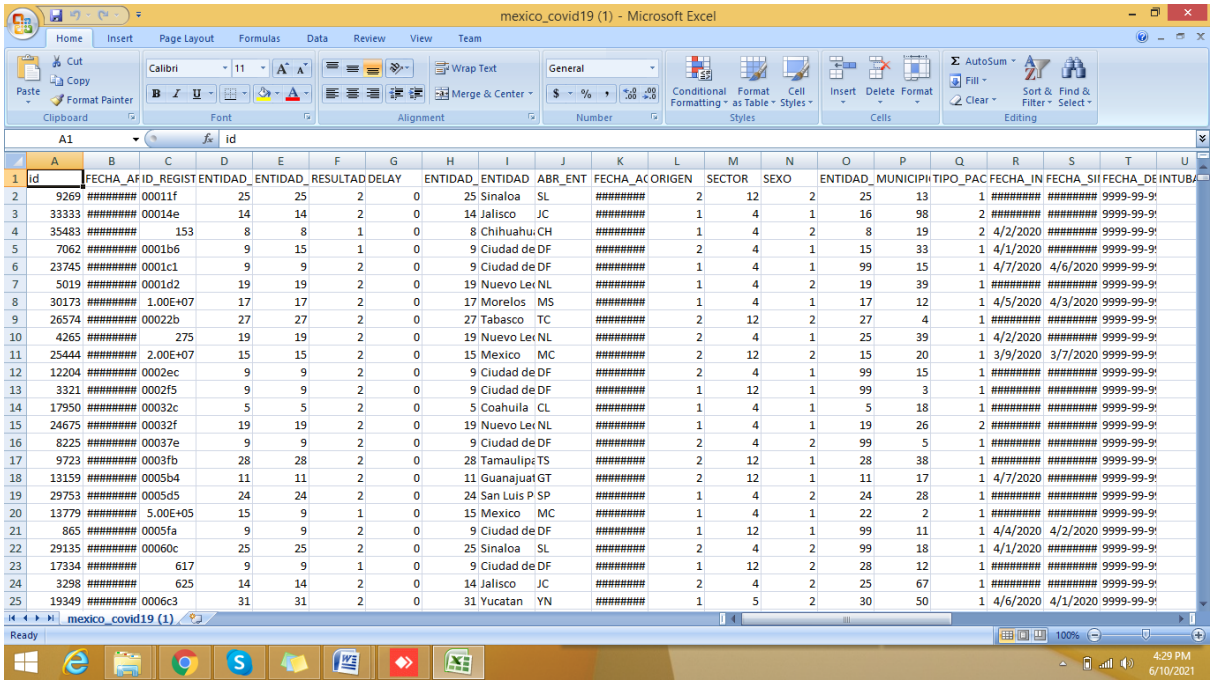
B. Data pre-processing: For applying machine learning techniques such that completeness can be introduced and a meaningful analysis can be achieved on the data, the data pre-processing is performed. This step delivers clean and denoised data for the feature selection process by removing redundant attributes from the dataset for enhancing the efficiency of the training model.

C. Feature selection: This step makes use of a subset comprising extremely unique features for diagnosing covid-19 diseases. These selective features relate to existing class of features. In the proposed method, the random forest model is applied for the feature selection. The random forest model takes 100 as the estimator value and generates tree structure of the most relevant features. RF classifier chooses those features which appear most appropriate or significant for predicting heart related disorders.

D. Classification: The mapping of chosen features is carried out to the training model for classifying provided features to make the prediction of disorder possible. A. Here, a kind of covid-19 disease is represented by each separate class. The logistic regression model is applied for the classification. The logistic regression takes input of the extracted features. In the research work, two classes are defined which are covid-19 and no covid-19.

VII. Result And Discussion

The supervised machine learning algorithms are applied for the covid-19 prediction. Python is used to run supervised machine learning algorithms including naive bayes, Bernoulli naive bayes, SVM, and voting classification. The Spyder editor of python is used and also required libraries like Sk-learn, numpyetc are used for the implementation. The dataset is of Mexico covid-19 which is collected from https://www.kaggle.com/marianarfranklin/mexico-covid19-clinical-data/metadata. The General Directorate of Epidemiology, Secretariat of Health of Mexico, submitted both positive and negative instances to be included in the dataset. The results of the RT-PCA test were used to create this dataset. About 30000 instances make up the dataset, and there are 31 characteristics. The first sort of feature in the dataset is one that is just necessary for hospitalization. Target collection, which includes positive and negative classes in the original dataset, will be changed to 0 and 1. For the Covid-19 forecast, zero means the worst-case scenario and 1 the best-case scenario. In the suggested model, technique of voting classification is employed for the COVID-19 forecast.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	FECHA_AF	ID_REGIST	ENTIDAD_	ENTIDAD_	RESULTAD	DELAY	ENTIDAD_	ENTIDAD	ABR_ENT	FECHA_AC	ORIGEN	SECTOR	SEXO	ENTIDAD_	MUNICIPI	TIPO_PAC	FECHA_IN	FECHA_SI	FECHA_DE	INTUBA
2	9269	########	00011f	25	25	2	0	25	Sinaloa	SL	########	2	12	2	25	13	1	########	########	9999-99-9	
3	33333	########	00014e	14	14	2	0	14	Jalisco	JC	########	1	4	1	16	98	2	########	########	9999-99-9	
4	35483	########	153	8	8	1	0	8	Chihuahu	CH	########	1	4	2	8	19	2	4/2/2020	########	9999-99-9	
5	7062	########	0001b6	9	15	1	0	9	Ciudad de	DF	########	2	4	1	15	33	1	4/1/2020	########	9999-99-9	
6	23745	########	0001c1	9	9	2	0	9	Ciudad de	DF	########	1	4	1	99	15	1	4/7/2020	4/6/2020	9999-99-9	
7	5019	########	0001d2	19	19	2	0	19	Nuevo Le	NL	########	1	4	2	19	39	1	########	########	9999-99-9	
8	30173	########	1.00E+07	17	17	2	0	17	Morelos	MS	########	1	4	1	17	12	1	4/5/2020	4/3/2020	9999-99-9	
9	26574	########	00022b	27	27	2	0	27	Tabasco	TC	########	2	12	2	27	4	1	########	########	9999-99-9	
10	4265	########	275	19	19	2	0	19	Nuevo Le	NL	########	2	4	1	25	39	1	4/2/2020	########	9999-99-9	
11	25444	########	2.00E+07	15	15	2	0	15	Mexico	MC	########	2	12	2	15	20	1	3/9/2020	3/7/2020	9999-99-9	
12	12204	########	0002ec	9	9	2	0	9	Ciudad de	DF	########	2	4	1	99	15	1	########	########	9999-99-9	
13	3321	########	0002f5	9	9	2	0	9	Ciudad de	DF	########	1	12	1	99	3	1	########	########	9999-99-9	
14	17950	########	00032c	5	5	2	0	5	Coahuila	CL	########	1	4	1	5	18	1	########	########	9999-99-9	
15	24675	########	00032f	19	19	2	0	19	Nuevo Le	NL	########	1	4	1	19	26	2	########	########	9999-99-9	
16	8225	########	00037e	9	9	2	0	9	Ciudad de	DF	########	2	4	2	99	5	1	########	########	9999-99-9	
17	9723	########	0003fb	28	28	2	0	28	Tamaulipa	TS	########	2	12	1	28	38	1	########	########	9999-99-9	
18	13159	########	0005b4	11	11	2	0	11	Guanajuat	GT	########	2	12	1	11	17	1	4/7/2020	########	9999-99-9	
19	29753	########	0005d5	24	24	2	0	24	San Luis P	SP	########	1	4	2	24	28	1	########	########	9999-99-9	
20	13779	########	5.00E+05	15	9	1	0	15	Mexico	MC	########	1	4	1	22	2	1	########	########	9999-99-9	
21	865	########	0005fa	9	9	2	0	9	Ciudad de	DF	########	1	12	1	99	11	1	4/4/2020	4/2/2020	9999-99-9	
22	29135	########	00060c	25	25	2	0	25	Sinaloa	SL	########	2	4	2	99	18	1	4/1/2020	########	9999-99-9	
23	17334	########	617	9	9	1	0	9	Ciudad de	DF	########	1	12	2	28	12	1	########	########	9999-99-9	
24	3298	########	625	14	14	2	0	14	Jalisco	JC	########	2	4	2	25	67	1	########	########	9999-99-9	
25	19349	########	0006c3	31	31	2	0	31	Yucatan	YN	########	1	5	2	30	50	1	4/6/2020	4/1/2020	9999-99-9	

Figure 2 Snapshot of Dataset

Performance Analysis Parameters

a. Accuracy: When compared to the total number of samples available for a program, accuracy is the proportion of samples whose categorization is performed correctly. This parameter's mathematical expression is expressed as follows:

$$A_i = \frac{t}{n} \cdot 100$$

Here, t represents the percentage of samples that are correctly identified, and n stands for the total case count for all samples.

b. Execution Time: It is the time interval between an algorithm's start and end points.

$$\text{Execution time} = \text{End time of algorithm- start of the algorithm}$$

c. Precision: The number of cases that are pronounced positive is divided by the total case count for all samples.

$$\textbf{Precision= TP/TP+FP}$$

d. Recall: By separating the real positive events from the total positive examples, recall is determined.

$$\textbf{Recall= TP/TP+FN}$$

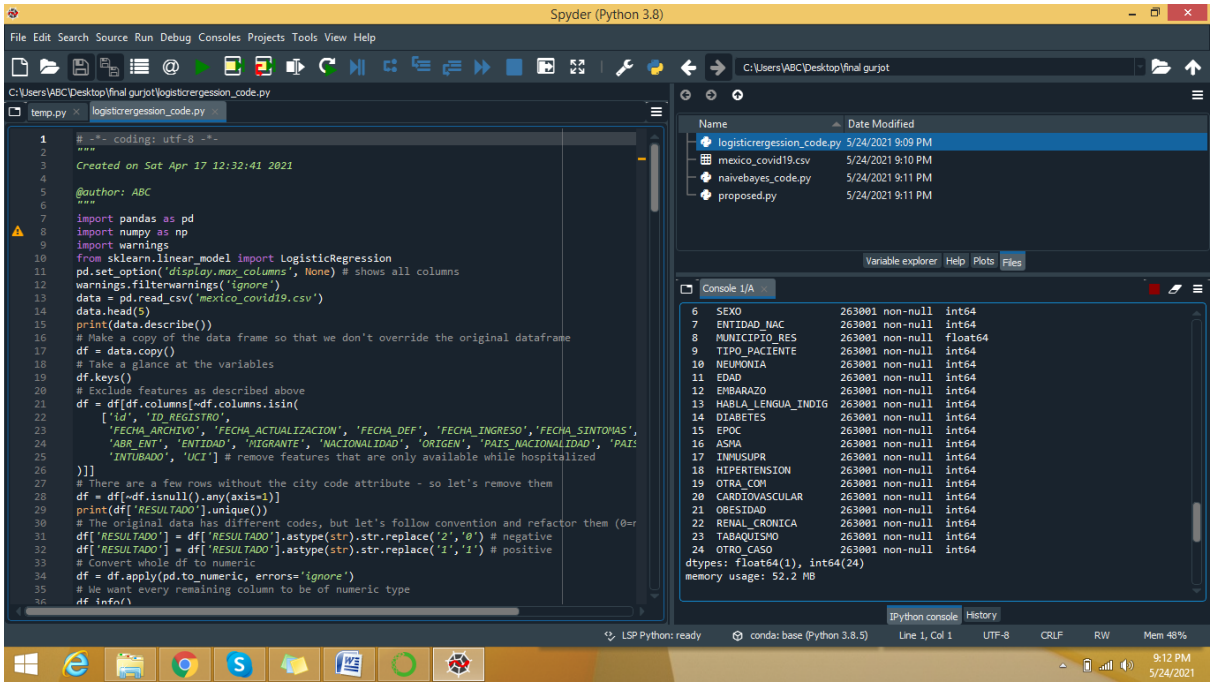


Figure 3: Interface of Anaconda

As shown in figure 3, the anaconda is the tool in which spyder is installed for the execution. The sypder interface is shown which processed for the execution of the python code. The console shows the final output of the project

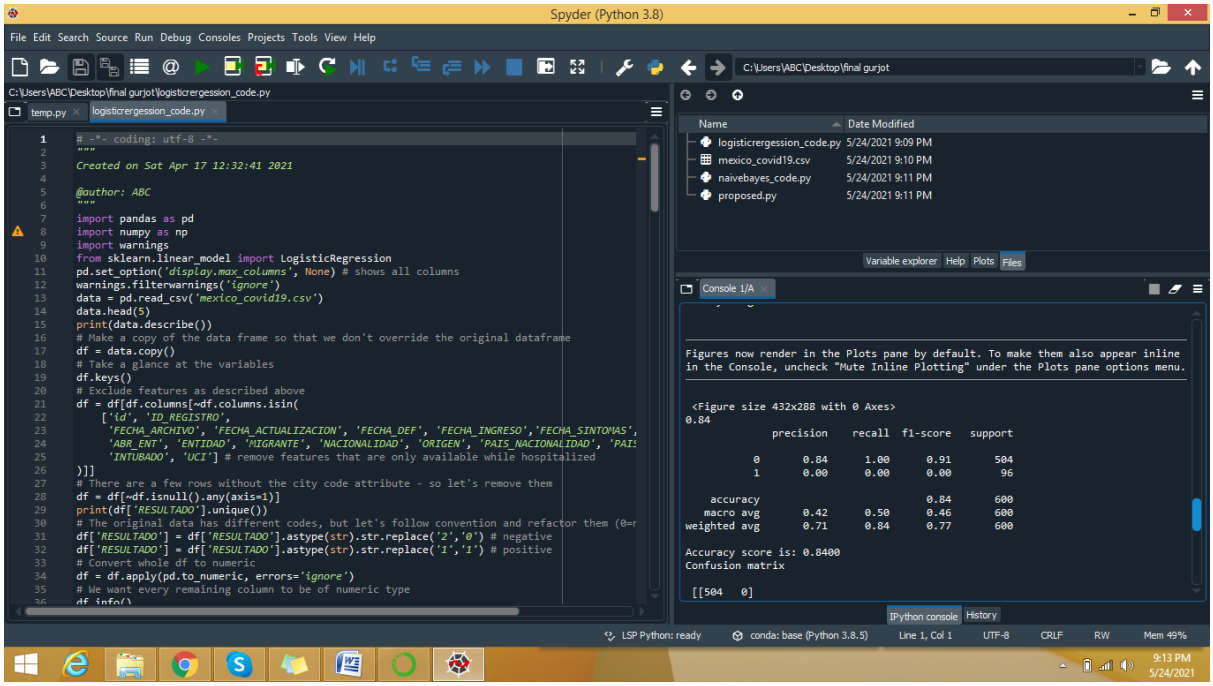


Figure 4 Apply logistic regression

The data that is obtained as input from the reliable source is pre-processed, as seen in figure 4. Testing and training data are separated. It is used to forecast the goal set for the COVID-19 using the logistic regression algorithm.

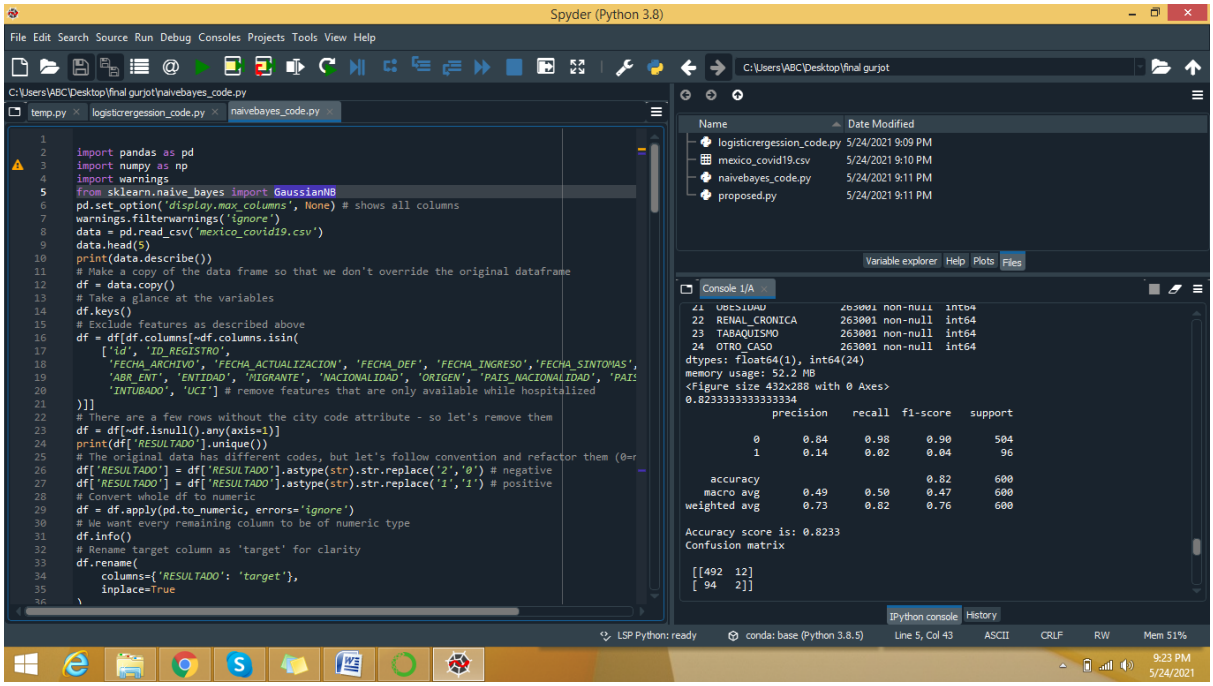


Figure 5: Apply Naive Bayes

The data that is retrieved as input from the reliable source has already undergone pre-processing, as seen in figure 5. Testing and training data are separated. The naive Bayes technique is used to estimate the goal set for the coronavirus.

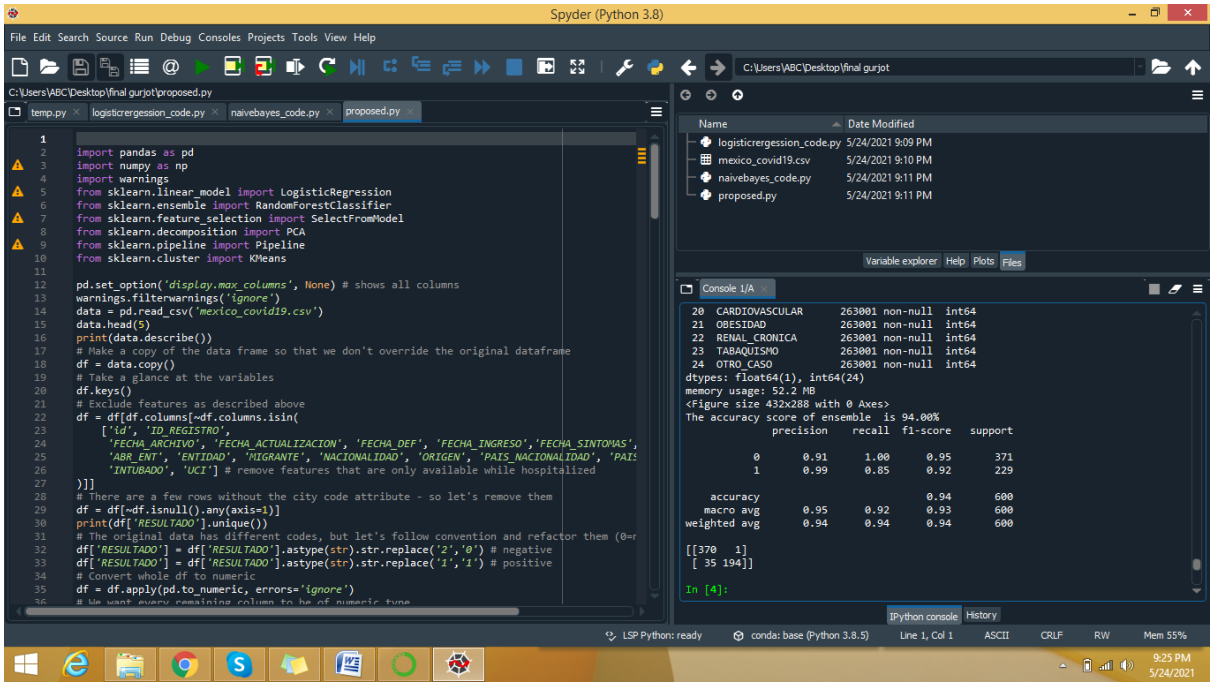


Figure 6: Apply Voting Classifier

The data that is retrieved as input from the reliable source has already undergone pre-processing, as seen in figure 6. Testing and training data are separated. The voting technique is used to estimate the goal set for the coronavirus.The Gaussian naive bayes, Bernoulli naive bayes, and random forest classifiers are all combined to create the voting classifier, which is used to forecast the Covid-19 outcome.

Table1: Performance Analysis

Parameters	Logistic Regression	Naive Bayes	Voting Classifier
Accuracy	84 percent	82 percent	94 percent
Recall	84 percent	82 percent	94 percent
Precision	71 percent	73 percent	94 percent

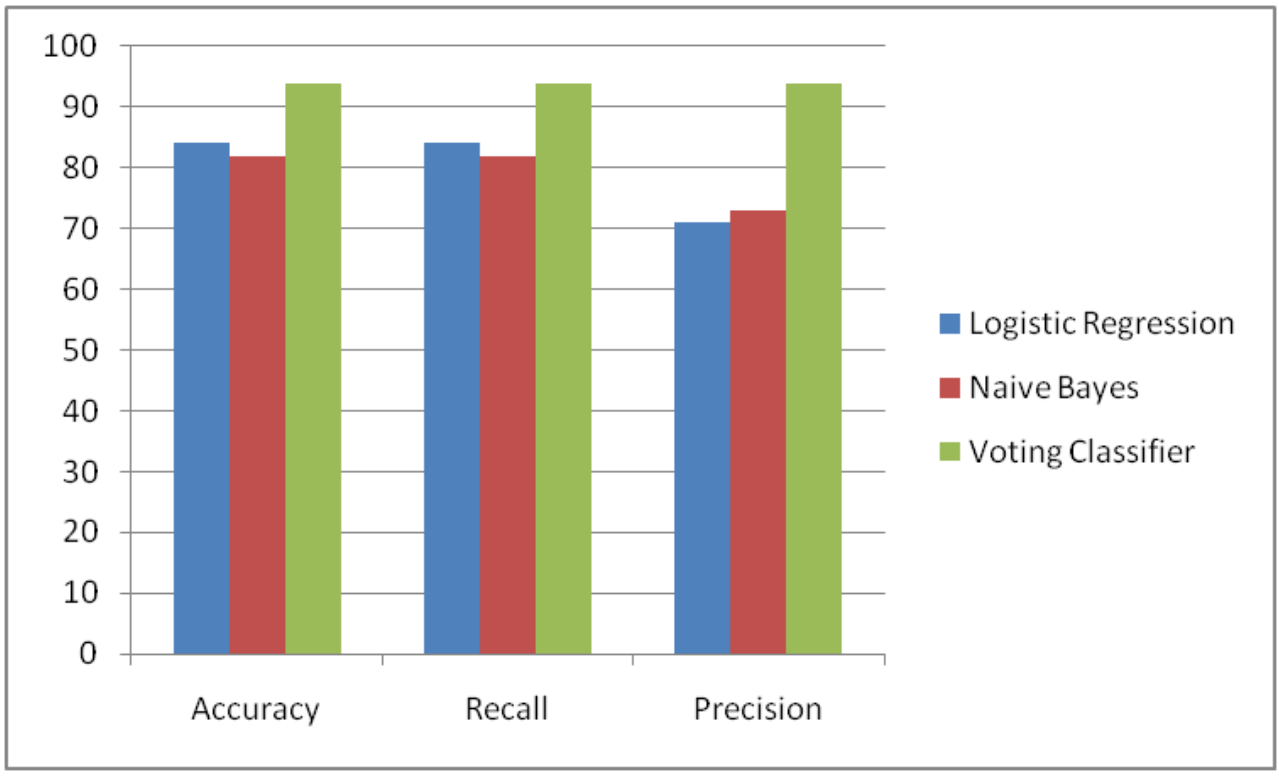


Figure 7 Graphically Analysis

Figure 7 compares the accuracy, precision, and recall of different classifiers, including logistic regression, naive bayes, and voting classifier. It is determined that voting classifiers perform the best for Covid-19 prediction in light of all three of these variables.

VIII. CONCLUSION

Early treatment of infectious diseases is crucial, and prevention and control are of utmost importance. First in the world to disclose a strange viral pneumonia was the Chinese province of Wuhan. The World Health Organization, the top global health organization, designated this virus as COVID-19, which it did on January 12, 2020. This virus is now the eighth coronavirus species identified to be able to infect people. The Covid-19 includes a wide variety of illnesses that can affect different organ parts. The analysis in this paper shows that Covid-19 prediction is highly difficult due to the numerous factors that it has for the coronavirus prediction, a variety of models, including decision trees, naive bayes, multilayer perceptron, and ensemble classifiers, are tested. To

predict Covid-19 diseases, a novel model integrating logistic regression and random forest is introduced. Utilizing RF, features are extracted, and logistic regression is used to perform the classification. The recall, accuracy and precision obtained from the proposed model is computed as 95 percent.

IX. REFERENCES

[1]. Ming-Yen Ng, Eric Yuk Fai Wan, Mary Sau-Man Ip, "Development and validation of risk prediction models for COVID-19 positivity in a hospital setting", 2020, International Journal of Infectious Diseases

[2]. Alberto Utrero-Rico, Javier Ruiz-Hornillos, Rocio Laguna-Goya, "IL-6–based mortality prediction model for COVID-19: Validation and update in multicenter and second wave cohorts", 2021, Journal of Allergy and Clinical Immunology

[3]. Arjun S Yadaw, Yan-chak Li, Gaurav Pandey, "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model", 2020, The Lancet Digital Health

[4]. Victor M. Castro, Chana A. Sacks, Thomas H. McCoy, "Development and External Validation of a Delirium Prediction Model for Hospitalized Patients With Coronavirus Disease 2019, 2021, Journal of the Academy of Consultation-Liaison Psychiatry

[5]. Ahmad Sedaghat, Shahab Band, Amir Mosavi, Laszlo Nadai, "COVID-19 (Coronavirus Disease) Outbreak Prediction Using a Susceptible-Exposed-Symptomatic Infected-Recovered-Super Spreaders-Asymptomatic Infected-Deceased-Critical (SEIR-PADC) Dynamic Model", 2020, IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)

[6]. Anwar Jarndal, Saddam Husain, Omar Zaatar, Talal Al Gumaei, Amar Hamadeh, "GPR and ANN based Prediction Models for COVID-19 Death Cases", 2020, International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)

[7]. D. Haritha, N. Swaroop, M. Mounika, "Prediction of COVID-19 Cases Using CNN with X-rays", 2020, 5th International Conference on Computing, Communication and Security (ICCCS)

[8]. Yifan Yang, Wenwu Yu, Duxin Chen, "Prediction of COVID-19 spread via LSTM and the deterministic SEIR model", 2020, 39th Chinese Control Conference (CCC)

[9]. Manoj Kumar, JatinBareja, Manjot Singh, Rupanshu Sharma, "An Intelligent Prediction Model of COVID-19 in India using Hybrid Epidemic Model", 2020, International Conference on Smart Electronics and Communication (ICOSEC)

[10]. OnderTutsoy, ŞuleÇolak, AdemPolat, Kemal Balikci, "A Novel Parametric Model for the Prediction and Analysis of the COVID-19 Casualties", 2020, IEEE Access

[11]. Nanning Zheng, Shaoyi Du, Jianji Wang, He Zhang, Wenting Cui, Zijian Kang, Tao Yang, Bin Lou, Yuting Chi, Hong Long, Mei Ma, Qi Yuan, Shupei Zhang, Dong Zhang, Feng Ye, Jingmin Xin, "Predicting COVID-19 in China Using Hybrid AI Model", 2020, IEEE Transactions on Cybernetics

[12]. Yuling Zou, "Prediction of the Initial Development Trend of COVID-19 with Dynamic Infectious Models", 2020, International Conference on Public Health and Data Science (ICPHDS)

[13]. ErtuğrulKaraçuha, NisaÖzgeÖnal, EsraErgün, Vasil Tabatadze, Hasan Alkaş, Kamil Karaçuha, HaciÖmerTontuş, Nguyen Vinh Ngoc Nu, "Modeling and Prediction of the Covid-19 Cases With Deep Assessment Methodology and Fractional Calculus", 2020, IEEE Access

[14]. Yixiao Ma, Zixuan Xu, Ziwei Wu, Yong Bai, "COVID-19 Spreading Prediction with Enhanced SEIR Model", 2020, International Conference on Artificial Intelligence and Computer Engineering (ICAICE)

[15]. Saud Shaikh, Jaini Gala, Aishita Jain, Sunny Advani, Sagar Jaidhara, Mani Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting", 2021, 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)

[16]. S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, F. Herrera, "COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images", 2020, IEEE Journal of Biomedical and Health Informatics

[17]. Md Masud Rana, Ahmed Abdelhadi, Md Riaz Uddin Ahmed, Ahad Ali, "Secure IoT Communication Systems for Prediction of COVID-19 Outbreak: An Optimal Signal Processing Algorithm", 2020, Third International Conference on Smart Systems and Inventive Technology (ICSSIT)

[18]. Narayana Darapaneni, Deepali Nikam, AnaghaLomate, Vaibhav Kherde, SwanandKatdare, Anwesh Reddy Paduri, Kameswara Rao, Anima Shukla, "Coronavirus Outburst Prediction in India using SEIRD, Logistic Regression and ARIMA Model", 2020, 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)

[19]. Ruizhi Han, Zhulin Liu, C. L. philip Chen, Lili Xu, Guangzhu Peng, "Mortality prediction for COVID-19 patients via Broad Learning System", 2020, 7th International Conference on Information, Cybernetics, and Computational Social Systems (ICCSS)

[20]. SinaArdabili, Amir Mosavi, Shahab S. Band, Annamaria R. Varkonyi-Koczy, "Coronavirus Disease (COVID-19) Global Prediction Using Hybrid Artificial Intelligence Method of ANN Trained with Grey Wolf Optimizer", 2020, IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE)

[21]. Kumar, Seba Susan, "COVID-19 Pandemic Prediction using Time Series Forecasting Models", 2020, 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)

[22]. Xiaoyi Fu, Xu Jiang, Yunfei Qi, Meng Xu, Yuhang Song, Jie Zhang, Xindong Wu, "An Event-Centric Prediction System for COVID-19", 2020, IEEE International Conference on Knowledge Graph (ICKG)

[23]. Yasin Khan, Pritam Khan, Sudhir Kumar, Jawar Singh, Rajesh M. Hegde, "Detection and Spread Prediction of COVID-19 from Chest X-ray Images using Convolutional Neural Network-Gaussian Mixture Model", 2020, IEEE 17th India Council International Conference (INDICON)

[24]. Andi Sulasikin, YudhistiraNugraha, Juan Kanggrawan, Alex L. Suherman, "Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta", 2020, IEEE International Smart Cities Conference (ISC2)

[25]. Chamara Sandeepa, CharukaMoremada, NadeekaDissanayaka, Tharindu Gamage, Madusanka Liyanage, "Social Interaction Tracking and Patient Prediction System for Potential COVID-19 Patients", 2020, IEEE 3rd 5G World Forum (5GWF)