STM JOURNALS

Journal of
Artificial Intelligence Research & Advances

ISSN: 2395-6720
Volume 8, Issue 2, 2021
DOI Journal: 10.37591/JoAIRA

http://computers.stmjournals.com/index.php?journal=JoAIRA&page=index

Research JoAIRA

Voting Based Classification Method for COVID-19 Prediction

Sushil Kumar*

Abstract

COVID-19 dataset comprises date, country, confirmed cases, recovered cases and total deaths. The data is integrated with climate data consisting of humidity, dew, ozone, perception, maximum temperature, minimum temperature, and UV. 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 various phases which include pre-processing, feature extraction, classification and performance analysis. In this research work, voting classification method is designed for the COVID-19 prediction. It is analyzed that proposed model increases accuracy, precision and recall for the COVID-19 prediction.

Keywords: COVID-19, machine learning, voting classification, feature extraction

INTRODUCTION

Though many countries have developed the vaccines, there is currently no exact treatment available to combat against novel corona virus. However, various symptoms can be treated, and treatment must be provided depending upon the medical condition of the patient. Furthermore, supplementary care for infectious people can contribute considerably [1, 2]. Maintain basic hand and respiratory hygiene, adhere to safe eating habits, and avoid close contact with anyone showing symptoms of respiratory disease (such as coughing or sneezing), etc. are some basic norms that one must follow for self-protection [3]. The extensive nature of COVID-19 forced factories to be shut down, schools to be suspended, people to be quarantined in their own homes, and thus considerably disrupted day-to-day life [4, 5]. Hence, reasonable prediction and analysis of the development tendency of this pandemic is the main key to get victory over it. Data mining refers to the analysis of data sets for finding interesting, new, and valuable patterns, relationships, models, and trends. The tasks of data mining include techniques based on artificial intelligence, machine learning, statistics, mathematics, and database systems. The data mining is mainly concerned with extracting information from a data set and transforming it into areas on able format for the use in future [6]. The COVID-19 prediction framework consists of six modules.

*Author for Correspondence
Sushil Kumar
E-mail: sk9630143@gmail.com

PG Scholar, Department of Computer Science & Engineering, Bansal Institute of Engineering & Technology, Lucknow, Uttar Pradesh, India

Received Date: May 17, 2021
Accepted Date: September 20, 2021
Published Date: October 20, 2021

Citation: Sushil Kumar. Voting Based Classification Method for Covid-19 Prediction. Journal of Artificial Intelligence Research & Advances. 2021; 8(2): 34–41p.

The modules are:

1. Data Collection module,
2. Pre-processing module,
3. Feature Selection,
4. Development of risk prediction module,
5. Prediction model validation and
6. Nomogram development and probability of COVID-19, as shown in Figure 1.

COVID-19 PREDICTION ALGORITHMS

Some popular COVID-19 prediction algorithms have been discussed below:

Multi-Layered Perceptron (MLP)

The biological nervous system has great influence on ANN (Artificial Neural Network) and it assist in processing the information in the same way as the brain [7, 8]. The basic component of this approach is a new structure of the information processing system. A number of highly interconnected processing components, known as neurons, compose the ANN system. These neurons perform together for tackling any issue. The learning process of this system is akin to humans using example.

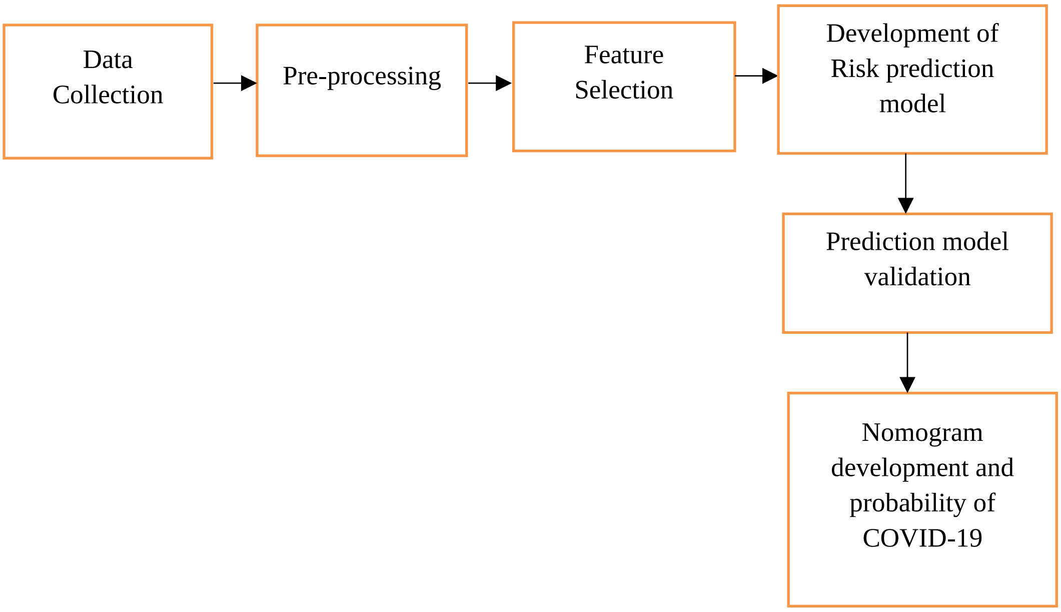
Data Collection
Pre-processing
Feature Selection
Development of Risk prediction model
Prediction model validation
Nomogram development and probability of COVID-19

Figure 1. COVID-19 risk prediction framework.

Support Vector Machine

SVM is a classic model that can be used not only for classification but also for regression [9, 10]. We do not delve into its theoretical derivation here; however,, the triangular fuzzy particles are selected and it membership function is expressed as:

$$A(x,a,m,b) = \begin{cases} 0, x < a \\ \frac{x-a}{m-a}, a \leq x \leq m \\ \frac{b-x}{b-m}, m < x \leq b \\ 0, x > b \end{cases} \tag{1}$$

LASSO

It is a regression algorithm based on LR (linear regression) method in which shrinkage is deployed. To shrink the extreme values of a data sample towards the central values is known as Shrinkage [11]. An ordinary multivariate regression makes the deployment of all the attributes present on it and a coefficient of regression is allocated to all. Therefore, the models become sparse with few coefficients during regularization as the process. The coefficients are removed in case of zero value [12]. This implies that the LASSO regression utilizes to minimize the following:

$$\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda \sum_{j-1}^{p}\left|\beta_j\right| \tag{2}$$

It sets the coefficient whose interpretation can be done as min (sum of square residuals +λ |slope|), in which, λ |slope| denotes the penalty term [13].

Linear Regression

The regression modelling includes the predication of a target class on the independent attributes [14]. Therefore, the association of independent variables with the dependent ones is discovered using

LR (Linear Regression). This method is utilized to carry out the prediction. LR is a kind of regression modelling and recognized as the most usable statistical method to accomplish the predictive analysis in ML (machine learning) [15, 16]. In LR, every observation can be done on the basis of two values such as dependent and independent variable. This technique is capable of determining a relationship between both the variables. While analyzing the linear regression, two factors (x, y) are considered [17–19]. The relation of y with x is called regression that is expressed in given equation as:

$$y = \beta_0 + \beta_1 + \varepsilon \qquad (3)$$

or equivalently,

$$E(y) = \beta_0 + \beta_1 x \qquad (4)$$

LITERATURE SURVEY

Ng *et al*. presented two validated risk prediction algorithms for COVID-19 positivity for which readily available parameters were considered in a general hospital setting [20–22]. 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 (positive predictive value) and NPV (negative predictive value). 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.

Utrero-Rico *et al*. designed a mortality prediction framework in order to predict the patients who were hospitalized due to COVID-19 [23]. 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.

Yadaw *et al*. intended an accurate predictive model of COVID-19 mortality in which unbiased computational techniques were utilized and the clinical attributes were also recognized [24]. The analysis related to development and validation of predictive model included the implementation of ML (machine learning) methods for clinical data which was taken from a huge cohort of patients who suffered from COVID-19 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.

Castro *et al*. suggested the supervised ML (machine learning) to EHR (electronic health record) data taken from three hospitals where the patients suffering from coronavirus disease 2019 were admitted [25]. Using this data, an incident delirium predictive framework was constructed. Those hospitals were considered for authenticating the framework. The c-index was found to be 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 of around 80%, its specificity was computed as 56% and negative predictive value was found to be 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.

Sedaghat *et al*. introduced a SEIR-PAD algorithm for assessing the susceptible, exposed, infected, recovered, super-spreader, and diseased populations [26]. There were seven sets of ordinary differential equations, having eight unknown coefficients, involved in this algorithm. 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.

Jarndal *et al*. developed a model for predicting the number of deaths occurred due to COVID-19 on the basis of documented number of older, diabetic and smoking cases [27]. This model was constructed using GPR (Gaussian Process Regression) technique. 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 COVID-19. Furthermore, this model was also assisted in preparing effective measures so that the number of deaths was reduced.

Haritha *et al*. recommended a TL (transfer learning) model in order to predict the COVID-19 from chest X-ray images [28]. The image was classified using GoogleNet that was an algorithm of CNN (Convolutional Neural Network). 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 to be 99% and accuracy in testing phase was computed to be 98.5%, while predicting the corona disease. In the remote places that had not any experienced practitioners, the primary health workers made the utilization of the recommended model.

Yang *et al.* projected the LSTM (Long Short-Term Memory) algorithm so that the infected population was predicted in China [29]. But, this algorithm was incapable of describing the dynamics of diffusion procedure and the error rate for long-term prediction was found greater. Thus, Susceptible-Exposed-Infected-Recovered (SEIR) was put forward later on for capturing the spread process of COVID-19. A sliding window technique was useful to estimate the parameter and predict the infected populations efficiently. The projected approach was useful for the epidemiological studies to understand the spread of the current COVID-19.

RESEARCH METHODOLOGY

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 process. These steps include pre-processing, feature extraction and classification. 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 analyzing the efficiency of ensemble classifier. The obtained outcomes indicate about the intricacy of this classifier which should be reduced for making the forecasting of COVID-19 possible.

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

Phases of COVID-19 Prediction

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. 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.

RESULT AND DISCUSSION

In this research work, the implementation and comparison of several models is performed for predicting the COVID-19 disease. The DT, Multilayer perception, NB, Ensemble classification method in which random forest, naïve Bayes models are combined, proposed models are compared with regard to certain performance parameters.

The Figure 2 illustrates that a variety of models including DT, NB, multilayer perceptron, ensemble and proposed models are compared concerning accuracy. The analytic results reveal that the proposed model achieves highest accuracy rate of almost 95% by performing better than other classifiers for predicting COVID-19.

As shown in Figure 3, the various models including DT, NB, MLP, ensemble and proposed models are compared in terms of precision. The analytic results reveal that the proposed model achieves highest precision rate of almost 95% by performing better than other classifiers for predicting COVID-19.

As shown in Figure 4, the various models like DT, NB, multilayer perceptron, ensemble are compared with the new model in terms of recall. It is analysed that recall of proposed model for COVID-19 prediction is approximately 95% which is higher than the other models.

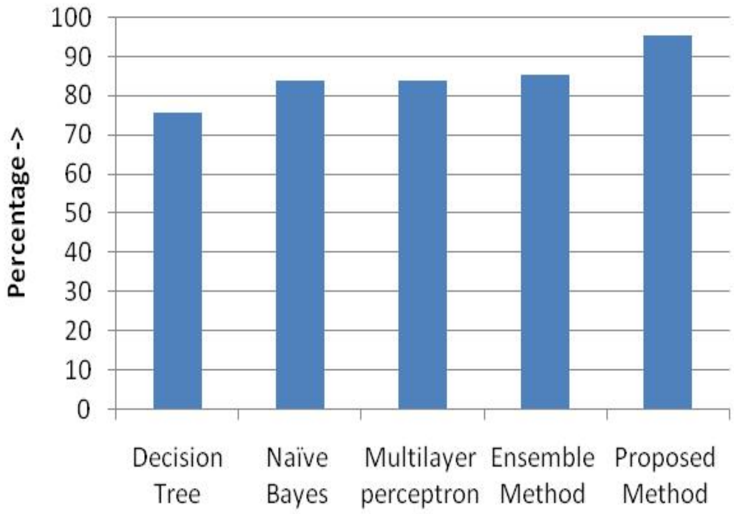

Figure 2. Accuracy analysis.

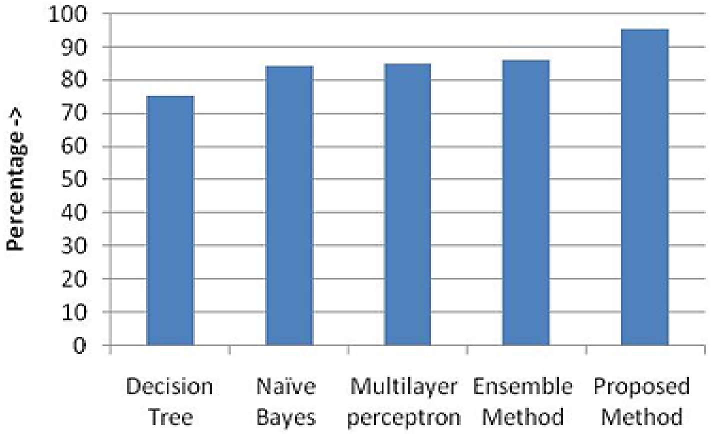

Figure 3. Precision analysis.

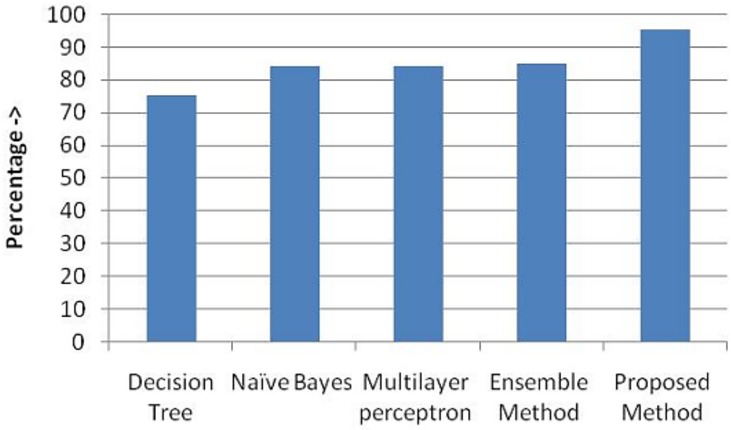

Figure 4. Recall analysis.

CONCLUSION

The COVID-19 consists of numerous kinds of diseases due to which various parts of the organs are infected. To conclude, it is analyzed in this work that COVID-19 prediction is very challenging as the large number of features included in it. The various models are tested for the COVID-19 prediction like decision tree, naïve Bayes, multilayer perceptron and ensemble classifier. The novel model in which the random forest and logistic regression are integrated is introduced to predict COVID-19 disorders. The extraction of features is generated using RF and the logistic regression is carried out to perform the classification. The recall, accuracy and precision obtained from the proposed model is computed as 95%.

REFERENCES

1. Zainab Abbas Abdulhussein Alwaeli, Abdullahi Abdu Ibrahim. Predicting Covid-19 Trajectory Using Machine Learning. 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2020 Oct 22–24; Istanbul, Turkey. New York: IEEE; 2020.
2. Suraj Bodapati, Harika Bandarupally, *et al.* COVID-19 Time Series Forecasting of Daily Cases, Deaths Caused and Recovered Cases using Long Short Term Memory Networks. 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA). 2020 Oct 30–31; Greater Noida, India. New York: IEEE; 2020.
3. Hanqing Chao, Xi Fang, *et al.* Integrative analysis for COVID-19 patient outcome prediction. Med Image Anal. 2020.
4. Danish Rafiq, Suhail Ahmad Suhail, *et al.* Evaluation and prediction of COVID-19 in India: A case study of worst hit states. Chaos Soliton Fract. 2020; 139: 110014.

5. Ardabili Sina F, Amir Mosavi, *et al*. COVID-19 Outbreak Prediction with Machine Learning. Algorithms. 2020; 13(10): 2–36.
6. Fotios Petropoulos, Spyros Makridakis. Forecasting the novel coronavirus COVID-19. PLOS ONE. 2020; 15(3): e0231236.
7. Zifeng Yang, Zhiqi Zeng, *et al*. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. 2020; 12(3): 165–174.
8. Pranay Nadella, Akshay Swaminathan, Subramanian SV. Forecasting efforts from prior epidemics and COVID-19 predictions. Eur J Epidemiol. 2020; 35: 727–729.
9. Giulia Giordano, Franco Blanchini, *et al*. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nat Med. 2020; 26: 855–860.
10. Jewell Nicholas P, Lewnard Joseph A, *et al*. Predictive Mathematical Models of the COVID-19 Pandemic Underlying Principles and Value of Projections. JAMA. 2020; 323(19): 1893–1894.
11. Saar Shoer, Tal Karady, *et al.* A Prediction Model to Prioritize Individuals for a SARS-CoV-2 Test Built from National Symptom Surveys. Clinical and Translational Resource and Technology Insights. 2021; 2(2): 196 -208.
12. Ferguson Neil M, Daniel Laydon, *et al*. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. London: Imperial College; 2020.
13. Rajan Gupta, Gaurav Pandey, *et al.* Machine Learning Models for Government to Predict COVID-19 Outbreak. Digital Government: Research and Practice (DGOV). 2020; 1(4): 1–6.
14. Ardabili Sina F, Amir Mosavi, *et al*. COVID-19 Outbreak Prediction with Machine Learning. Algorithms. 2020; 13(10): 249.
15. Challener Douglas W, Dowdy Sean C, *et al*. Analytics and Prediction Modeling during the COVID-19 Pandemic. Perspective and Controversy. 2020; 95(9): S8–S10.
16. Collins Gary S, Maarten van Smeden, Rile Richard D. COVID-19 prediction models should adhere to methodological and reporting standards. Eur Respir J. 2020; 56: 2002643.
17. Lara Jehi, Xinge Ji, *et al*. Individualizing Risk Prediction for Positive Coronavirus Disease 2019 Testing: Results from 11,672 Patients. Chest. 2020; 158(4): 1364–1375.
18. Dong Ji, Dawei Zhang, *et al.* Prediction for Progression Risk in Patients with COVID19 Pneumonia: The CALL Score. Clin Infect Dis. 2020; 71(6): 1393–1399.
19. Yue Xiang, Yonghong Jia, *et al.* COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models. Infect Dis Model. 2021; 6: 324–342.
20. Tarik Alafif, Reem Alotaibi, *et al*. On the prediction of isolation, release, and decease states for COVID-19 patients: A case study in South Korea. ISA Trans. In Press, Corrected Proof. ScienceDirect. 2021.
21. Salam Bennouar, Abdelghani Bachir Cherif, *et al*. Development and validation of a laboratory risk score for the early prediction of COVID-19 severity and in-hospital mortality. Intensive Crit Care Nurs. 2021; 64: 103012.
22. Ming-Yen Ng, Eric Yuk Fai Wan, *et al.* Development and validation of risk prediction models for COVID-19 positivity in a hospital setting. Int J Infect Dis. 2020; 101: 74–82.
23. Alberto Utrero-Rico, Javier Ruiz-Hornillos, *et al.* IL-6–based mortality prediction model for COVID-19: Validation and update in multicenter and second wave cohorts. Allergy Clin Immunol. 2021; 147(5): 1652 -1661.
24. Yadaw Arjun S, Yan-Chak Li, *et al*. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health. 2020; 2(10): e516–e525.
25. Castro Victor M, Sacks Chana A, *et al*. Development and External Validation of a Delirium Prediction Model for Hospitalized Patients with Coronavirus Disease 2019. J Acad Consult-Liaison Psychiatry. 2021; 62(3): 298–308.
26. Ahmad Sedaghat, Shahab Band, *et al*. 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). 2020 Nov 18–19; Budapest, Hungary. New York: IEEE; 2021.

27. Anwar Jarndal, Saddam Husain, *et al*. GPR and ANN based Prediction Models for COVID-19 Death Cases. 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI); 2020 Nov 3–5; Sharjah, United Arab Emirates. New York: IEEE; 2020.
28. Haritha D, Swaroop N, Mounika M. Prediction of COVID-19 Cases Using CNN with X-rays. 2020 5th International Conference on Computing, Communication and Security (ICCCS); 2020 Oct 14–16; Patna, India. New York: IEEE; 2020.
29. Yifan Yang, Wenwu Yu, Duxin Chen. Prediction of COVID-19 spread via LSTM and the deterministic SEIR model. 2020 39th Chinese Control Conference (CCC); 2020 Jul 27–29; Shenyang, China. New York: IEEE; 2020.