Multimedia Tools and Applications (2024) 83:82177–82198
https://doi.org/10.1007/s11042-024-18317-6

Computer aided disease detection and prediction of novel corona virus disease using machine learning

S. M. Saravanakumar[1] · **T. Revathi**[1]

Received: 27 June 2023 / Revised: 9 November 2023 / Accepted: 19 January 2024 /
Published online: 12 March 2024

Abstract
Machine Learning is recent emerging technique in prediction of various health related issues in medical system. It is very essential to predict the COVID-19 virus before it spreads and affects an entire community. Machine Learning is being used to detect the presence of COVID-19 virus as early as possible by analyzing patient's health condition and collecting data such as gender, age, Body Mass Index (BMI), asthma symptoms, wheezing, dyspnea, respiratory failure, cough, blood sugar level etc., with this information used eighteen machine learning algorithms such as ELM, Logistic Regression, SGD, KNN, SVM, QDA, LDA, XGBoost etc., to analyze the data and predict the presence of COVID-19 virus. Table and Charts are plotted with the help of the results acquired from the machine learning algorithm. As a result, early prediction of COVID-19 becomes possible and huge loss in terms of both health and economy can be avoided.

Keywords Machine Learning · Computer Aided Disease Diagnosis · Corona Virus Disease · COVID19 · Lung Disease Diagnosis

1 Introduction

In 2019, the novel coronavirus disease (COVID-19) ravaged the world. As of July 23, 2021, there are about 192 million infected people worldwide, and 4.1365 million deaths. At present, the new coronavirus is still spreading and circulating in many places around the world, especially the emergence of the Delta variant has increased the risk of COVID-19 becoming a pandemic again. The symptoms of COVID-19 are diverse.Fever, dry cough and exhaustion are the most common mild symptoms in patients. In between 15.7% and 32.0% of individuals, serious symptoms will appear. After a week, severe cases frequently experience dyspnea, and they develop multiple organ failure, acute respiratory distress syndrome, septic shock, difficult-to-treat metabolic acidosis, and coagulation dysfunction. Studies have indicated that severely unwell patients have a very bad prognosis and have a much higher fatality rate. Therefore, early detection of patients with high risk of disease progression and provision of intensive

✉ S. M. Saravanakumar
saravmath@gmail.com

[1] Department of Computer Science, PSG College of Arts & Science, Coimbatore, India

care and appropriate intervention, such as optimizing respiratory support and management of critically ill patients with new coronary pneumonia, will help improve the success rate of critically ill patients and rationally allocate medical resources. The risk factors for severe disease in patients with new coronary pneumonia include advanced age, comorbidity with other diseases, and elevated levels of D-dimer (D-Dimer), etc. At present, there is still no effective drug for the treatment of new coronary pneumonia. Early detection of patients at risk of developing serious illnesses is crucial for the prevention and treatment of epidemics.

Machine Learning (ML) is the core of artificial intelligence, and its data processing, induction, and synthesis capabilities are far superior to other statistical methods. Machine learning methods have obvious advantages in clinical outcome prediction and risk factor assessment. In this study, on the basis of using machine learning to predict the severe trend of new coronary pneumonia in the early stage, combined with a large number of clinical characteristics of COVID-19 patients, prediction model was constructed to effectively predict the risk of severe disease of patients and evaluate the accuracy of the model. It provides important technical support for clinical decision-making and a reliable prediction model for the prevention and treatment of new coronary pneumonia. Building an accurate and effective forecasting model for major infectious diseases based on multi-machine learning can predict outbreak trends and help formulate countermeasures in advance. This study uses mathematical modeling to predict the trend of the number of people diagnosed with the recent outbreak of new coronavirus (COVID-19) pneumonia through machine learning under limited data. According to the information released by relevant departments, the epidemic situation is predicted. The time when the inflection point appeared, and compared the proportion of the estimated final confirmed cases in each province, based on this, roughly divided the severity of the epidemic, which has guiding significance for the protection work of people in various provinces and cities. The clinical indicators and outcomes (in-hospital death and in-hospital tracheal resection) of patients with COVID-19 admitted to hospitals were collected from February 5 to April 15, 2020, using Artificial Neural Network (ANN), and machine learning algorithms to construct a predictive model for the patient's clinical outcome.

First, 21 indicators with significant differences were selected as the input features of the training model, optimization was performed on the constructed model to adjust parameters, and the optimal feature combination was screened based on feature importance; further analysis of the value of each feature for the positive and negative impact of the prediction results, to quantify and attribute the importance of each feature; evaluate the performance of the prediction model, compare it with other machine learning methods, and discuss its advantages. The prediction performance of different machine learning has its own strengths, and the ensemble prediction model effectively integrate the advantages of multiple machine learning, so as to obtain stable and accurate prediction results.

This paper is organized into 5 sections. Section 1 starts with the introduction, Section 2 deals with the literature survey related to the paper, Section 3 imparts research methods, Section 4 insists on discussion about the topic and finally Section 5 draws the conclusion which is followed by references reviewed for this paper.

2 Literature survey

The Susceptible-Infected-Removed (SIR) model and the Susceptible-Exposed-Infected-Removed (SEIR) model are classic models for the study of infectious disease dynamics. Ren Lei et al. used the fractional SIR model [1] to predict the rollout of COVID-19. Fan

Ruguo et al. predicted the inflection point of the epidemic under three different periods based on the epidemic SEIR dynamics model [2]. Alenezi et al. used the SIR model [3] to evaluate and foretell the rollout of COVID in Kuwait. The SIR model they proposed was almost This coincides with actual confirmed cases of infection and recovered cases. However, each outbreak has its own unique transmission characteristics, prevention and control policies in different regions, and population movement will increase the uncertainty of the infectious disease dynamics model. Therefore, traditional infectious disease models are difficult to provide reliable results for long-term predictions. With the advancement of deep learning technology, extreme learning methods have shown good performance in the field of prediction. Compared with other methods, deep learning has the advantages of fast calculation speed, small error and more accurate prediction results. Based on existing COVID-19 data sets, some researchers have used deep learning models to foretell the spread of COVID-19. Chimmula and Zhang used the LSTM neural network [4] state-of-the-art machine learning model to predict the likely end time of the pandemic in Canada. Based on their LSTM model, their model was 93.4% accurate in the short term and 92.67% in the long term, estimating the time it would take to end the pandemic at about three months.

Ismail et al. based on Autoregressive Integrated Moving Average (ARIMA) model, LSTM, Nonlinear Autoregressive Neural Network (Nonlinear Auto Regression Neural Network, NARNN) model [5] conducted a comparative study for predicting COVID-19 cases in Denmark, Belgium, Germany, France, United Kingdom, Finland, Switzerland, and Turkey. They found that LSTM provided The Smallest Root Mean Square Error (RMSE) compared to other models. Mehdi et al. used Recursive Neural Network (RNN), LSTM, Seasonal Autoregressive Integrated Moving Average (SARIMA) and Holt-Winter's exponential smoothing and moving average methods [6] To predict new coronary pneumonia cases in Iran, they compared these methods and found that the LSTM model was better than other models in predicting infection in Iran. Parul et al. used Deep LSTM, Convolutional LSTM, and BiLSTM models [7] to predict the number of positive cases of new coronary pneumonia in India. Their research results showed that the error of the BiLSTM model in short-term predictions (1–3 days) was less than 3%, which was very accurate. forecast result. Nahla et al. proposed to use LSTM and Gate Recurrent Unit (GRU) model [8] to predict confirmed cases and deaths in Egypt, Saudi Arabia and Kuwait. Their research results showed that the confirmed cases of LSTM in the three countries Best in cases and GRU in fatal cases in Egypt and Kuwait. Verma, H et al. designed recurrent and convolutional neural network models: vanilla LSTM, stacked LSTM, ED_LSTM, BiLSTM, CNN and CNN-LSTM models [9], predicting the daily confirmed cases in India and its four most affected areas on 7, 14, and 21 days. Through their research, they found that the BiLSTM and CNN-LSTM models have better prediction results than other models. However, neither these traditional models nor deep learning models take into account the impact of factors such as vaccination, population size, and medical facilities on the spread of the new COVID epidemic.

In response to the epidemic, a large number of scholars collected epidemic data from the Wuhan Municipal Health Commission (hereinafter referred to as the Health Commission), and carried out a series of research on epidemic prediction and evaluation of epidemic prevention measures. Fan Ruguo et al. used the infectious disease model SEIR [11–13] to simulate the peak confirmed cases in Wuhan corresponding to different incubation periods, and predicted that the inflection point of the epidemic would appear from February 20 to 25, 2020 [16]. Yang et al. [14] evaluated the government's control measures and made an authoritative prediction on the development trend of the epidemic by

amending SEIR combined with the Artificial Intelligence (AI) method of training. Wang et al. [15] combined the infectious disease model SIR (susceptible infectious recovered) with machine learning methods to assess the severity of the epidemic in important domestic provinces and cities, and predicted the final number of confirmed cases. Zhang Lin [16] based on the GGM (Gordon Growth Model) model, divided into three stages of barrier-free exponential growth, exponential growth and sub-linear growth to fit the model with a high degree of agreement with the epidemic report data, revealing the spread mechanism of COVID-19 and the epidemic was predicted. Huang et al. [17] ARIMA model was used to predict the seasonal hand, foot and mouth disease epidemic in Jiangmen City. To sum up the above studies, the SEIR mechanism model and GGM and other phenomenological models can play a good role in data theory support and scientific control and prevention guidance in the prediction of this epidemic, but the parameters are not considered comprehensively, and the total number of model populations is constant or asymptotically constant. The basic assumptions are inconsistent with the dynamic changes of epidemic parameters at different stages [18]; the network dynamics model plays an auxiliary role in epidemic prediction and verification through data learning, and can refine and improve the model. However, the data in the introductory stage of the outbreak is not complete, limited and different, and data learning is limited. Current research mainly uses LSTM ARIMA algorithms to simulate groups. Individual differences were not considered [19].

Agent can operate independently in a certain environment, acting on its own generation environment is also affected by the external environment, and can continuously acquire knowledge from the environment to improve its own ability [20]. Agent-based modeling and simulation methods is the main method of modeling individual intelligent behavior [21, 22]. Compared with the dynamic infectious disease model, the infectious disease model of the agent has the most important feature that it can model and simulate the spreading process of the infectious disease from the individual level, and describe the microcosm of the spreading individual of the epidemic through the interaction between Agent entities. The behavior and macro epidemic situation are more in line with the realistic transmission mechanism of infectious diseases. Chen Bin et al. [10] used the multi-agent simulation method to evaluate the epidemic prevention measures of COVID-19, and concluded that "early isolation" is still the efficient way to handle epidemic, and the epidemic is initially suppressed. There will be "partial bubbling" events, and it is necessary to prepare for a protracted war.

3 Research methodology

Various types of methodology have been carried out by collect different kinds of data from the COVID patient. Condition assessment, data processing, quality control and model establishment are done accordingly.

3.1 Collect general clinical indicators

The general clinical index questionnaire was designed by the researchers based on literature review and clinical experience, including: (1) General information: gender, age, stage of disease course (stable stage, acute exacerbation stage), Body Mass Index (BMI), education level. (2) Disease factors: asthma symptoms (yes, no), wheezing (yes, no), dyspnea (yes, no), loss of appetite (yes, no), cough (yes, no), cough description, sputum description,

description of sputum characteristics, time of peak symptoms, times of acute attacks, time from last acute attack, pathogenic factors, hospitalization history (yes, no) and times of acute attacks, outpatient visits for acute attacks (yes, no), systemic use of corticosteroids (yes, no), combined with respiratory failure (yes, no), combined with pulmonary heart disease (yes, no), combined with pulmonary encephalopathy (yes, no), abnormal nutrition metabolism (yes, no), cardiovascular disease (yes, no)), other disease history (yes, no), COPD family history (yes, no), smoking history (yes, no), oxygen inhalation (yes, no), daily oxygen inhalation time, oxygen inhalation flow rate, oxygen inhalation method, Use of transcutaneous oxygen saturation monitor (yes, no), blood oxygen saturation value, use of non-invasive ventilation (yes, no), daily non-invasive ventilation time, non-invasive ventilation method, wearing mask (yes, no), whether aware of non-invasive ventilation How to disinfect the ventilator humidification tank and ventilator tubing, nutritional status, exercise (yes, no), exercise method, daily exercise time, pursed-lip abdominal breathing (yes, no), use of inhalants (yes, no), inhalation Type and quantity of doses, long-term use of inhaled drugs (yes, no).

3.2 Condition assessment of covid patients

COVID was assessed using the COVID Assessment Test (CAT) [6]. CAT includes 8 aspects: cough, expectoration, chest tightness, shortness of breath, limited activity, confidence to go out, sleep, and energy. The total score of the CAT score is 40 points. The total score >30 is considered to be very serious. The condition is considered to be severe if the score is $20<$ total score ≤ 30. The condition is moderate if the score is $10<$ total score ≤ 20. Spend.

3.3 Assessment of the severity of dyspnea in covid patients

Dyspnea severity in COVID patients was assessed using the Dyspnea Index Score (mMRC) [6]. mMRC is a modified British MRC, with a total of 0–4 grades. Grade 0: only experienced during physically demanding activity; Grade 1: having trouble breathing while moving swiftly on flat ground; Grade 2: moving slowly than peers when on flat ground, or stopping to rest when out of breath; Grade 3 walking on flat ground for many minutes or roughly 100 m before having to stop and pant; Grade 4: Dyspnea while putting on and taking off clothes or being unable to leave the house due to acute dyspnea.1.1.4 Pulmonary function tests and determination of severe airflow limitation.

The lung function and ventilation of the subjects were measured by the German Jaeger (Master Screen PFT System) lung function instrument after calibration of environmental parameters, volume calibration and gas calibration.

Referring to the GOLD (2019 edition) guideline [5], this study used the percentage of forced expiratory volume in 1 s (FEV 1%) in the pulmonary function index to determine the range of airflow limitation in COVID patients, and mild airflow limitation is FEV $1\% \geq 80\%$; moderate airflow limitation is $50\% \leq$ FEV $1\% < 80\%$; severe airflow limited range is $30\% \leq$ FEV $1\% < 50\%$; extreme airflow limitation is FEV $1\% < 30\%$.

In order to establish a risk model, with FEV $1\% = 50\%$ as the critical threshold, patients were divided into those at risk of severe airflow limitation (severe, very severe FEV $1\% < 50\%$) assigned a value of 1, and those without risk of severe airflow limitation (mild, moderate FEV $1\% \geq 50\%$) is assigned a value of 0.

3.4 Quality control

A scientific research team composed of 1 postgraduate student, 1 pulmonary function technician and 4 undergraduate nurses with more than 5 years of work experience in respiratory department who passed the unified training and assessment conducted questionnaire design, questionnaire survey and pulmonary function test. The questionnaire uses uniform instructions and is filled out anonymously by patients. If the patient is unable to fill in the questionnaire by himself, the investigator will read the questionnaire items without any inducement, and then let the patient make independent judgments on the result of item Lung function tests were performed with the same equipment and the same technicians. After the survey was completed, the questionnaires were returned on the spot. Data entry, desensitization, and analysis are not performed by the same personnel. Two investigators entered the questionnaire results, and the other member desensitized the data before handing it over to the data analyst. In this study, 20 cases of COVID patients were pre-investigated, and the problems existing in the implementation of the pre-investigation were revised and improved. Once complete, the formal survey and data collection begins.

3.5 Data processing

In this study the public dataset from the Huoshenshan Hospital (HSSLL024) was included. According to the diagnostic criteria of COVID-19, a total of 143 patients were included in this study. 100 patients were randomly selected as the training set, and the data of the remaining 43 patients were used as the test set. The degree of severe and above airflow limitation (FEV $1\% < 50\%$) was used as the output variable of the model, and the remaining 50 variables were used as input variables for data processing and model research. Review the variable characteristics of the data, and conduct preliminary screening according to the established rules: (1) delete variables with a missing data ratio of $> 90\%$ in each column; (2) delete variables with a single category ratio $> 90\%$ in each column; (3) delete Column variables with coefficient of variation (variable coefficient, CV) < 0.05. This data set has missing values and the missing values are filled using KNN imputation. In this study, three feature screening methods, namely no screening, Lasso screening and Boruta screening, were used to extract the features of the dataset. Use Lasso screening and Boruta screening to generate important feature data, which reflect the importance of each input variable to the result prediction. After 4 kinds of missing value processing and 3 kinds of feature screening methods, a total of 12 processed data sets were obtained.

3.6 Model establishment

Use the train_test_split package in the Python Scikit-Learn library to divide the data into 80% training set and 20% test set. The training set data is used for model training, and the test set data is used for evaluating and selecting models. The 12 preprocessed datasets were modeled separately using 17 kinds of machine learning and 1 kind of ensemble learning algorithm (Ensemble Learning). 17 machine learning algorithms including: Logistic Regression, Stochastic Gradient Descent (SGD), K Nearest Neighbors (KNN), Decision Tree, Gaussian Naïve Bayes, Bernoulli Naive Bayesian (Bernoulli

Naive Bayes), Multinomial Naive Bayes (Multinomial Naive Bayes), Support Vector Machine (SVM), Quadratic Discriminant Analysis (QDA), Random Forest (Random Forest), Extreme Random Tree (Extra Tree), Linear Discriminant Analysis (LDA), Passive Aggressive, AdaBoost, Bagging, Gradient Boosting, XGBoost, Extreme Learning Machine (ELM). The results of the ensemble models are voted by the best top 5 models.

3.7 Experimental results

The results of the experiments follow; different kinds of analysis have been carried out using various parameters with the help of extreme machine learning algorithm.

3.8 Model evaluation

The Area Under the ROC Curve (AUC), accuracy rate, precision rate, recall rate, and F1 value are used as indicators for model evaluation. When the results of each indicator are inconsistent, AUC is used as the main reference. On the training set, model evaluation was performed using ten-fold cross-validation. In the test set, Bootstrapping algorithm is used to resample 200 times for external verification. The evaluation index of the test set data is used as the basis for the best model selection.

3.9 Sample size verification

Use the screened best model, randomly use 10%, 20%…100% of the data in the training set to train the model, and use the test set data to evaluate the prediction performance of the trained model. The method is repeated 100 times to observe the effect of changes in the training sample size on the model's predictive performance. Model building and graph visualization use Python3.7.3 + Pycharm to build a development environment, and use Scikit-Learn library and Xgboost library to build a machine learning model.

3.10 Statistical methods

Data were analyzed using MATLAB software. The measurement data is represented by ($\pm$s). In the comparison of different data preprocessing methods, if the data is distributed normally and the difference is comparable, the comparison between multiple groups is performed by analysis of variance; if the data is not normally distributed or the variance is not uniform, multiple groups The Kruskal-Walli's test was used for comparison. Count data are expressed as frequency and percentage. $P < 0.05$ was considered statistically significant.

3.11 General clinical indicators of the research subjects

A total of 432 sampling was disseminated in the study, and 418 sampling was recovered, with an effective recovery rate of 96.7%. Among the 418 COVID patients included, there were 46 females and 372 males; age (63.7 ± 10.9) years; 304 cases in the stable stage and 114 cases in the acute exacerbation stage; 206 cases (49.3%) had mild or moderate airflow limitation, there were 212 cases (50.7%) of severe and extremely severe cases. A total of 50 input variables and 1 output variable were collected, and the variables are shown in Table 1.

Table 1 General Information of the Included COVID Patients ($n=418$)

variable	data		variable	data	
Age (±s, years old)	63.7 ± 10.9		The number of days since the last acute attack outpatient visit a (±s,d)	0.6 ± 1.8	
Gender [n (%)]	Female	46 (11.0)	Systemic hormone use [n (%)]	none	403(96.4)
	male	372(89.0)		Have	15 (3.6)
Disease course stage [n (%)]	stable period	304(72.7)	Cor pulmonale [n (%)]	none	407 (97.4)
	Acute exacerbation	114(27.3)		Have	11 (2.6)
BMI(±*s*, kg/m^2)	23.1 ± 3.6				
Education level a [n (%)]	illiteracy		Abnormal nutrition metabolism [n (%)]	none	416 (99.5)
	primary school	150 (36.0)		Have	2 (0.5)
	junior high school	145 (34.8)	Cardiovascular disease [n (%)]	none	408 (97.6)
	High School/Technical Secondary School	55 (13.2)		Have	10 (2.4)
	College and above	42 (10.0)			
			History of other diseases [n (%)]		
Asthma Symptoms [n (%)]	none	79 (18.9)	Asthma symptoms [n (%)]	none	300 (71.8)
	Have	339(81.1)		Have	118 (28.2)
Wheezing [n (%)]	none	82 (19.6)	COPD family history [n (%)]	none	260 (62.2)
	Have	336 (80.4)		Have	158 (37.8)
Dyspnea [n (%)]	none	62 (14.8)	Smoking History [n (%)]	none	91 (21.8)
	Have	356 (85.2)		Have	327 (78.2)

Table 1 (continued)

variable	data		variable	data	
mMRC grade a [n (%)]	0 class	25 (6.0)	Oxygen Inhalation [n (%)]	none	389 (93.1)
	1 class	145 (34.8)			
	2 class	178 (42.7)		Have	29 (6.9)
	3 class	68 (16.3)			
	4 class	1 (0.2)			
Loss of appetite [n (%)]	none	358 (85.6)	Using a transcutaneous oximeter [n (%)]	none	413 (98.8)
	Have	60 (14.4)		Have	5 (1.2)
Cough [n (%)]	none	71 (17.0)	Exercise [n (%)]	none	109 (26.1)
	Have	347 (83.0)		Have	309 (73.9)
The number of acute attacks (±s, times)	1.4 ± 1.5		Pursed-lip abdominal breathing [n (%)]	none	256 (61.2)
				Have	162 (38.8)
Days from last acute attack (±s, d)	1.4 ± 31.7		CAT score (±s, points)	12.8 ± 5.6	

Table 1 (continued)

variable	data		variable	data	
Pathogenic factor b [n (%)]	don't know	139 (33.4)	Use of inhalants [n (%)]	none	47 (11.2)
	cold	244 (58.7)		Have	371 (88.8)
	cold air	8 (1.9)			
	other	8 (1.9)			
	sports	6 (1.4)	Long-term use of inhaled drugs [n (%)]	none	56 (13.4)
	irritating gas	11 (2.7)			
Acute attack hospitalization times (±s, times)	0.6 ± 1.1			Have	361 (86.6)
			FEV1% in lung function [n (%)]	≥ 50%	206(49.3)
				< 50%	212(50.7)

A means missing 1 case ($n = 417$); b means missing 2 cases ($n = 416$); *BMI* = Body Mass Index, *mMRC* = dyspnea index score, *COPD* = Chronic Obstructive Pulmonary disease, *FEV1%* = Forced Expiratory Volume in 1 s as a percentage of predicted value, *CAT* = Chronic Obstructive Lung Disease Assessment Form

Table 2 Total Data Set Variable Elimination

Variable Name	Variable Name
Oxygen inhalation (yes, no) ❷	nutritional status ❸
Daily oxygen inhalation time ❷	Blood oxygen saturation value❸
Oxygen flow ❷	Use of non-invasive ventilation (yes, no) ❷
Oxygen inhalation method ❷	non-invasive ventilation time per day❸
Non-invasive ventilation method ❷	Wear a face covering (yes, no)❷
humidification of non-invasive ❷	Using transcutaneous oximetry
How to Disinfect Tanks and Ventilator Tubing ❷	Meter (yes, no) ❷

❶Variables with a missing data ratio of > 90% in each column; ❷the proportion of a single category in each column > 90% of the variables; ❸Variables with a coefficient of variation of each column < 0.05

3.12 Data review and preliminary screening results

According to the above data review and preliminary screening principles, 12 input variables were eliminated. The reasons for variable elimination are summarized in Table 2.

3.13 Key factors affecting airflow restriction

A total of twelve processed data sets and ranking along with twelve factors that affects airflow limitations were obtained after evaluating 4 different types of value treatment along with three types of feature screening. From this study, mMRC grade along with age, body mass index, smoking history, computed aided tomography and dyspnea were the first in the variable characteristics ranking in addition to this these are vital indicators for models demonstrating the importance of the role. Results are obtained after incorporating non-filling screening method and lesso screening methods as shown in Fig. 1. The top three predictors are mMRC grade, smoking history and dyspnea, about 54.15% of feature importance acquired when it comes to mMRC grade. The results obtainedby using unpacking and volta screening methods is shown in Fig. 2. The three most crucial predictors are computed aided tomography score, age and mMRC, with computed axial tomography score comes to 26.64%.

3.14 Early warning model establishment and evaluation

To model twelve datasets, we are using one ensemble learning algorithm and eighteen different machine learning algorithms. All the eighteen algorithms were compared for the accuracy and as a result a sum of 216 models was obtained. The results of eighteen machine learning algorithms are shown in Table 3. There was a significant different in the statistics when comparing the performance of different algorithms (i.e.) ($P < 0.05$), the highest value (0.738 ± 0.089) is obtained from average ACU of stochastic gradient descent algorithm. Using the Bootstrapping algorithm, the test set was validated externally, the results are shown below in table. In the similar way the predictive ability of models was also compared using different algorithms and the difference was remarkable ($P < 0.05$), the largest value of (0.757 ± 0.057) was obtained from the average AUC of the integrated learning algorithm. Bootstrap algorithm was used to assess the performance of 4 different missing value treatments and 3 types of trait examinations. The

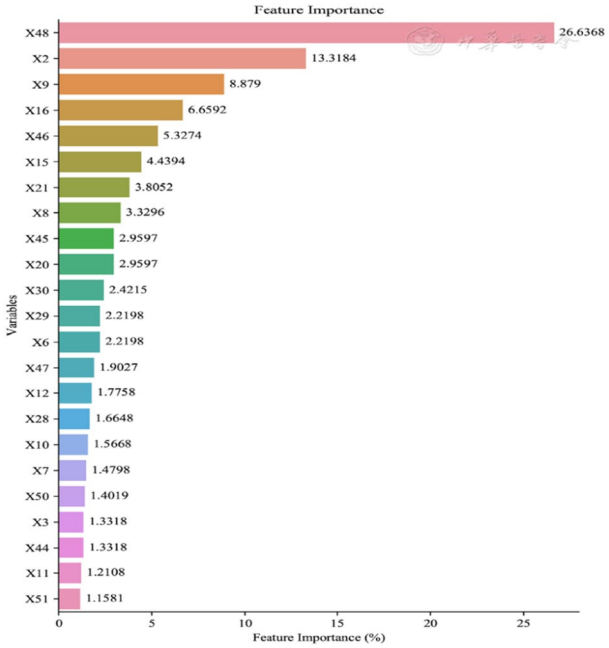

Fig. 1 Unfilled Boruta screened feature importance maps

results are tabulated in Tables 4 and 5 respectively. The efficiency of the model can be increased by avoiding padding and Lesso filtering and the difference is appreciable ($P < 0.05$).

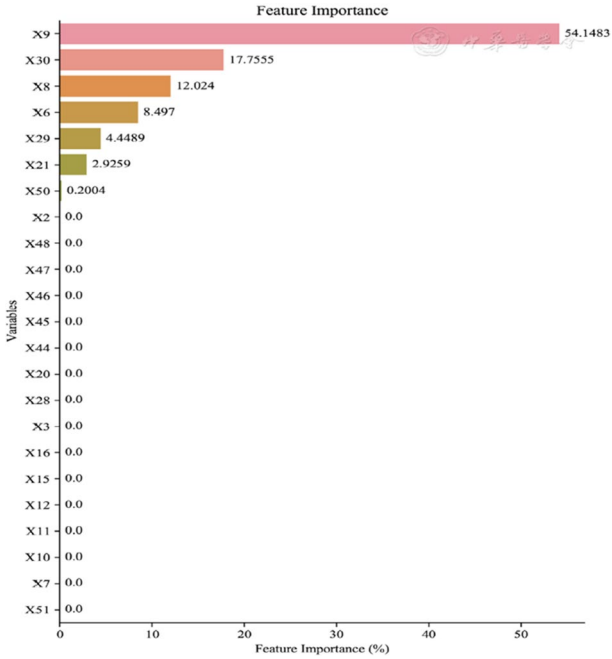

Fig. 2 Unfilled Lasso screened feature importance maps

Table 3 Sky 10-foldValidation Results of 18 Machine Learning Algorithms

Machine Learning Algorithm	AUC		Accuracy		Precision Rate		Recall Rate		F1 Value	
	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI
AdaBoost	0.707±0.099	(0.688, 0.723)	0.683±0.076	(0.670,0.697)	0.672±0.073	(0.659, 0.685)	0.740±0.106	(0.721, 0.759)	0.701±0.079	(0.687, 0.715)
Bagging	0.666±0.101	(0.648, 0.684)	0.626±0.087	(0.611, 0.642)	0.642±0.096	(0.624, 0.659)	0.614±0.121	(0.592, 0.636)	0.622±0.091	(0.605, 0.638)
Bernoulli Naive Bayes	0.714±0.073	(0.701, 0.728)	0.664±0.061	(0.653, 0.675)	0.659±0.067	(0.647, 0.672)	0.709±0.093	(0.693, 0.726)	0.680±0.063	(0.668, 0.691)
Decision Tree	0.693±0.073	(0.679, 0.706)	0.685±0.071	(0.672, 0.697)	0.667±0.065	(0.656, 0.679)	0.758±0.115	(0.737, 0.779)	0.706±0.076	(0.692, 0.720)
Extra Tree	0.679±0.081	(0.664, 0.693)	0.664±0.071	(0.651,0.676)	0.664±0.074	(0.651, 0.678)	0.694±0.110	(0.674, 0.714)	0.674±0.074	(0.661, 0.687)
Gaussian Naive Bayes	0.703±0.085	(0.688, 0.718)	0.639±0.067	(0.627, 0.651)	0.616±0.058	(0.605, 0.626)	0.777±0.079	(0.763, 0.791)	0.685±0.058	(0.675, 0.696)
Gradient Boosting	0.699±0.096	(0.681, 0.716)	0.664±0.082	(0.649, 0.678)	0.662±0.079	(0.647, 0.676)	0.695±0.131	(0.671, 0.719)	0.673±0.091	(0.656, 0.689)
KNN	0.696±0.087	(0.680, 0.712)	0.637±0.082	(0.622, 0.652)	0.618±0.076	(0.605, 0.632)	0.781±0.120	(0.760, 0.803)	0.684±0.072	(0.671, 0.697)
LDA	0.730±0.092	(0.713, 0.746)	0.677±0.072	(0.664, 0.690)	0.676±0.075	(0.662, 0.689)	0.704±0.111	(0.684, 0.724)	0.685±0.077	(0.671, 0.699)
Logistic Regression	0.729±0.095	(0.712, 0.746)	0.682±0.074	(0.669, 0.696)	0.683±0.077	(0.669, 0.697)	0.701±0.117	(0.680, 0.722)	0.687±0.084	(0.672, 0.703)
Multinomial Naive Bayes	0.639±0.099	(0.621, 0.658)	0.596±0.089	(0.580, 0.612)	0.590±0.087	(0.574, 0.606)	0.697±0.141	(0.672, 0.723)	0.632±0.089	(0.616, 0.648)
Passive Aggressive	0.648±0.112	(0.627, 0.668)	0.601±0.090	(0.584, 0.617)	0.603±0.102	(0.585, 0.622)	0.639±0.184	(0.606, 0.672)	0.607±0.124	(0.584, 0.629)
QDA	0.720±0.090	(0.704, 0.736)	0.661±0.076	(0.647, 0.674)	0.650±0.074	(0.637, 0.664)	0.723±0.114	(0.703, 0.744)	0.681±0.078	(0.667, 0.695)
Random Forest	0.665±0.111	(0.645, 0.685)	0.625±0.099	(0.607, 0.643)	0.636±0.107	(0.616, 0.655)	0.620±0.131	(0.597, 0.644)	0.623±0.108	(0.603, 0.642)
SGD	0.737±0.088	(0.721, 0.754)	0.685±0.075	(0.672, 0.699)	0.684±0.077	(0.670, 0.698)	0.716±0.110	(0.696, 0.736)	0.695±0.077	(0.681, 0.709)
SVM	0.721±0.102	(0.700, 0.737)	0.666±0.087	(0.651, 0.682)	0.678±0.098	(0.660, 0.695)	0.666±0.112	(0.645, 0.686)	0.667±0.090	(0.651, 0.683)
XGBoost	0.676±0.098	(0.660, 0.696)	0.637±0.079	(0.622, 0.651)	0.642±0.078	(0.628, 0.656)	0.642±0.124	(0.620, 0.665)	0.637±0.090	(0.621, 0.654)
ELM	0.747±0.093	(0.721, 0.752)	0.702±0.102	(0.686, 0.721)	0.697±0.092	(0.684, 0.712)	0.625± 0.125	(0.580, 0.630)	0.612± 0.097	(0.615, 0.647)
p-value	<0.000 1		<0.000 1		<0.000 1		<0.000 1		<0.000 1	

Table 4 External Verification Results of 18 Machine Learning Algorithms

machine learning algorithm	AUC		Accuracy		Precision		Recall rate		F1 value	
	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI
AdaBoost	0.716±0.068	(0.713, 0.718)	0.678±0.061	(0.676, 0.680)	0.659±0.073	(0.656, 0.662)	0.765±0.070	(0.763, 0.768)	0.706±0.060	(0.704, 0.708)
Bagging	0.657±0.074	(0.654, 0.660)	0.627±0.059	(0.624, 0.629)	0.623±0.076	(0.620, 0.626)	0.670±0.086	(0.667, 0.673)	0.643±0.067	(0.640, 0.645)
Bernoulli Naive Bayes	0.697±0.062	(0.694, 0.699)	0.648±0.057	(0.646, 0.650)	0.639±0.073	(0.636, 0.642)	0.721±0.069	(0.718, 0.724)	0.675±0.058	(0.673, 0.677)
Decision Tree	0.681±0.065	(0.678, 0.684)	0.683±0.061	(0.681, 0.686)	0.658±0.074	(0.655, 0.661)	0.799±0.069	(0.797, 0.802)	0.719±0.058	(0.717, 0.721)
Ensemble Learning	0.757±0.057	(0.755, 0.760)	0.708±0.056	(0.706, 0.711)	0.695±0.074	(0.692, 0.698)	0.771±0.074	(0.768, 0.774)	0.728±0.057	(0.725, 0.730)
Extra Tree	0.666±0.065	(0.664, 0.669)	0.658±0.062	(0.655, 0.660)	0.646±0.077	(0.643, 0.649)	0.733±0.089	(0.729, 0.737)	0.683±0.064	(0.680, 0.685)
Gaussian Naive Bayes	0.654±0.066	(0.651, 0.656)	0.610±0.057	(0.608, 0.612)	0.597±0.070	(0.595, 0.600)	0.728±0.074	(0.725, 0.731)	0.654±0.060	(0.651, 0.656)
Gradient Boosting	0.707±0.064	(0.705, 0.710)	0.655±0.065	(0.653, 0.658)	0.645±0.079	(0.642, 0.648)	0.726±0.074	(0.723, 0.729)	0.680±0.065	(0.678, 0.683)
KNN	0.663±0.071	(0.660, 0.666)	0.633±0.066	(0.630, 0.636)	0.609±0.080	(0.606, 0.612)	0.809±0.087	(0.806, 0.813)	0.690±0.060	(0.688, 0.693)
LDA	0.714±0.060	(0.712, 0.716)	0.678±0.053	(0.676, 0.680)	0.665±0.070	(0.662, 0.667)	0.743±0.070	(0.740, 0.746)	0.699±0.056	(0.697, 0.701)
Logistic Regression	0.721±0.062	(0.718, 0.723)	0.689±0.056	(0.687, 0.692)	0.678±0.072	(0.675, 0.681)	0.748±0.069	(0.746, 0.751)	0.709±0.058	(0.707, 0.711)
Multinomial Naive Bayes	0.651±0.064	(0.648, 0.654)	0.602±0.068	(0.600, 0.605)	0.602±0.081	(0.598, 0.605)	0.668±0.122	(0.663, 0.673)	0.627±0.080	(0.624, 0.630)
Passive Aggressive	0.686±0.075	(0.683, 0.689)	0.624±0.082	(0.621, 0.628)	0.636±0.095	(0.632, 0.639)	0.626±0.200	(0.618, 0.634)	0.613±0.126	(0.608, 0.619)
QDA	0.686±0.067	(0.683, 0.688)	0.646±0.061	(0.643, 0.648)	0.630±0.074	(0.627, 0.633)	0.753±0.075	(0.750, 0.756)	0.683±0.061	(0.681, 0.686)
Random Forest	0.687±0.066	(0.685, 0.690)	0.659±0.063	(0.657, 0.662)	0.657±0.076	(0.654, 0.660)	0.692±0.088	(0.689, 0.696)	0.671±0.069	(0.668, 0.674)
SGD	0.718±0.064	(0.715, 0.720)	0.672±0.054	(0.670, 0.674)	0.657±0.071	(0.655, 0.660)	0.747±0.075	(0.744, 0.750)	0.697±0.058	(0.694, 0.699)
SVM	0.708±0.061	(0.705, 0.710)	0.648±0.072	(0.645, 0.650)	0.641±0.083	(0.637, 0.644)	0.709±0.082	(0.706, 0.712)	0.671±0.072	(0.668, 0.674)
XGBoost	0.680±0.069	(0.677, 0.683)	0.639±0.066	(0.637, 0.642)	0.636±0.082	(0.632, 0.639)	0.697±0.081	(0.694, 0.700)	0.662±0.067	(0.659, 0.664)
ELM	0.724±0.071	(0.721, 0.727)	0.699±0.068	(0.693, 0.698)	0.686±0.075	(0.691, 0.699)	0.663±0.072	(0.686, 0.693)	0.653±0.071	(0.648, 0.653)
p-value	<0.000 1		<0.000 1		<0.000 1		<0.000 1		<0.000 1	

Table 5 The Results of External Evaluation of Missing Value Processing Methods

Approach	AUC		Accuracy		Precision		Recall rate		F1 value	
	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI
Not	0.724±0.070	(0.723, 0.725)	0.689±0.070	(0.688, 0.691)	0.676±0.083	(0.674, 0.677)	0.751±0.098	(0.749, 0.753)	0.707±0.073	(0.705, 0.708)
Random Forest	0.682±0.068	(0.681, 0.684)	0.640±0.063	(0.638, 0.641)	0.630±0.074	(0.628, 0.631)	0.722±0.101	(0.720, 0.724)	0.668±0.072	(0.667, 0.669)
Random Forest Improve	0.681±0.069	(0.680, 0.683)	0.642±0.063	(0.640, 0.643)	0.632±0.076	(0.631, 0.634)	0.720±0.101	(0.718, 0.722)	0.669±0.073	(0.667, 0.670)
Simple	0.679±0.068	(0.677, 0.680)	0.642±0.064	(0.641, 0.644)	0.634±0.079	(0.633, 0.636)	0.720±0.104	(0.718, 0.722)	0.669±0.073	(0.668, 0.671)
p-value	<0.000 1		<0.000 1		<0.000 1		<0.000 1		<0.000 1	

Table 6 Results of External Evaluation of Different Feature Screening Methods

Screening method	AUC		Accuracy		Precision		Recall rate		F1Score	
	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI	(±s)	95%CI
Boruta filtering	0.681±0.072	(0.680, 0.682)	0.652±0.068	(0.650, 0.653)	0.643±0.081	(0.641, 0.644)	0.722±0.100	(0.721, 0.724)	0.676±0.073	(0.674, 0.677)
Lasso filtering	0.703±0.069	(0.701, 0.704)	0.651±0.069	(0.649, 0.652)	0.643±0.082	(0.642, 0.644)	0.717±0.110	(0.715, 0.719)	0.672±0.079	(0.671, 0.674)
No filtering	0.691±0.071	(0.690, 0.692)	0.658±0.068	(0.656, 0.659)	0.643±0.078	(0.642, 0.645)	0.745±0.094	(0.743, 0.746)	0.687±0.071	(0.686, 0.688)
*P*value	<0.000 1		<0.000 1		0.534 4		<0.000 1		<0.000 1	

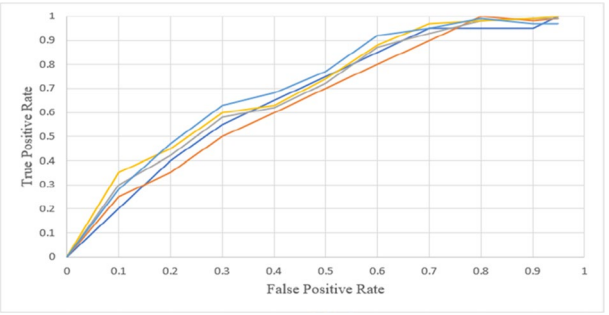

Fig. 3 ROC Curves for Optimal Risk Prediction Models

S*election of early warning model.*

Use the test data to test 216 machine learning models and select the model with the highest AUC as the best model. Table 6 shows the best five AUC prediction performance indicators of the model. The AUC is 0.7909, the precision rate is 75.90%, the precision rate is 75.00%, the recall ratio is 78.57%, and the F1 value is 0.767. See Figs. 2 and 3 for ROC and PR curves.

3.15 Sample size verification

Select the algorithm corresponding to the best model as the algorithm for sample size verification. The data set is divided into training set and test set in the ratio 8:2. 10%, 20%…100% of the training set samples were randomly selected for model training, and the process was repeated 100 times. The results show that when the sample size reaches about 70%, the curve tends to be flat. Tip at this point, the sample size will no longer increase the prediction performance (Figs. 1, 2, 3 and 4).

4 Discussion

Because of the increase in severity of airflow limitation the risk of death from COVID increases. Hence, it is important to define the value of airflow limitation and execute suitable intervention methods [7]. This study predicts the severity of airflow limitation in

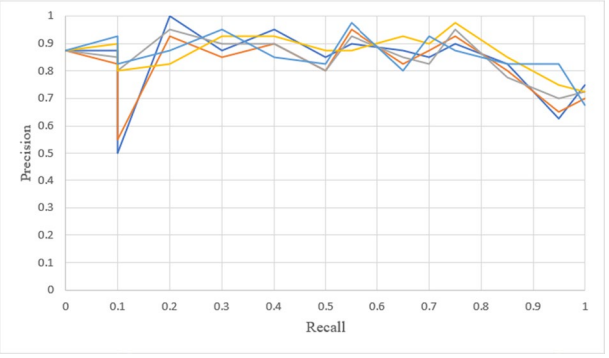

Fig. 4 P-R Curves of Optimal Risk Prediction Model

COVID patients by constructing a risk model of severe airflow limitation. Through the data mining process such as preliminary screening, missing value filling, and variable feature screening, the integrated learning model is selected using AUC, accuracy, precision, recall, and F1 value as evaluation indicators for internal verification, the best model, and external verification. for the best model. The results of this study are consistent with those of LIU et al. [8]. Integrated learning [9] achieves better predictive performance by combining multiple learning algorithms. The joint decision-making of multiple learning algorithms is more accurate than the prediction using a single learning algorithm, which has certain clinical application value. Dong Quanming et al. [10] established a FEV 1 early warning model using a multiple linear model, but this study did not include disease-related factors for discussion. Based on previous studies, this study comprehensively considered the variables of disease-related factors to construct a classification model for the degree of airflow limitation. ZAFARI et al. [11] developed an individualized prediction model for lung function decline, but this study only included smokers with mild to moderate COPD, and it cannot be predicted for patients with severe COPD. In addition, several studies are all predictions of the absolute value of lung function FEV 1, while FEV 1% is a relatively individualized evaluation index, which is more widely used in clinical practice and has attracted more attention from researchers, so it is more predictive value [12, 13].

The key parameters to build the model for this study are mMRC grade, age, BMI, CAT score, smoking history, and dyspnea. Pulmonary ventilation function indicators are related to age and smoking history, which is consistent with previous studies [14–16]. COPD patients are mostly elderly. With age, the contractility of respiratory muscles decreases, the elastic recoil of the thorax and lungs decreases, the bronchial tube wall shrinks, and the lumen narrows, resulting in increased pulmonary ventilation resistance and slowed airflow. In addition, the patient's smoking affects the lung microecological group and reduces the lung defense ability [17]. Lung function is related to BMI, which is consistent with some scholars' research [18, 19]. GRIGSBY et al. [20] showed that in developing countries, the lower the BMI, the worse the lung function. However, some scholars pointed out that FEV 1 has nothing to do with BMI, only positively correlated with height [10, 14]. The discrepancies in the study results may be due to the differences in the sociodemographic characteristics of the included populations. Lung function was also significantly correlated with mMRC grade and CAT score [5, 21, 22]. Predictions based on patient lung function have shifted to an assessment of disease exacerbation risk and symptoms that can be used to refine COPD airflow limitation classification.

Missing data has become a common and unavoidable serious problem in real-world research. In data analysis, if the entire information of the patient is deleted due to a small part of data missing, most of the information will be lost; if too much information is lost and features are added, it may increase noise instead and affect the final result. In this study, the variables whose missing data accounted for more than 90% were initially screened out and then filled. As for whether the remaining data needs to be filled, and how to fill it effectively, until today, there has not been a consensus. The most commonly used imputation method is to input with the mean, median, or data with the highest frequency, but the accuracy is also low. STEKHOVEN et al. [23] established the random forest iterative imputation method, and achieved a good filling effect. Random forest can effectively handle mixed types of data filling, which has advantages over single-type filling methods. However, the results of this study showed that different filling methods had statistically significant effects on model performance ($P < 0.05$). When the data is not filled, the effect of the early warning model is better. The reason for this result is that the non-filling method in this study is different from the non-filling

method in previous studies. This is a method that maximizes the original data set for analysis method, so the best effect is obtained.

No screening, Lasso screening and Boruta screening are used in this study to examine variable functions. Various screening methods can reduce the number of features, reduce dimensionality, reduce learning difficulty, improve model performance, and improve model generalization ability. The feature screening methods implemented in the study affects the efficiency of the model. Among them, none of the screenings include all the variables after manipulating missing values to understand the predictive power of the model. However, if only a less characteristics are considered to create the model, the time duration can be reduced to a great extent and the accountability of the model can be improved. In Boruta screening, all sets of characteristics correlated with the dependent variable are selected so that the influencing factors of the dependent variable can be understood. Compared to ordinary least squares estimation, Lasso filtering can quickly and efficiently extract important variables to simplify the model when there are many variables. In univariate analysis for hypothesis testing in this study, Lasso screening performed better in the model with an average AUC value of (0.719 ± 0.09), but Lasso screening did not show good predictive performance among the top five models. Ensemble learning, without padding, Boruta sieve model outperforms ensemble learning, without padding, Lasso sieve model. ELM model performed with higher accuracy and precision when compared to other models (Tables 3 and 7) and the results are finalized based on it.

The innovations of this study: (1) There is no mature risk warning model for severe airflow limitation in COVID patients. The machine learning model established in this study provides an auxiliary decision-making basis for disease assessment of COVID patients. (2) So far, many studies on machine learning often use one or several machine learning algorithms to build models, and rarely use different data preprocessing methods for diversified modeling to compare model prediction performance. However, this study adopted as many as 216 algorithms through different data cleaning methods, and established 2 160 models through ten-fold cross-validation. At the same time, this study uses the advanced Bootstrapping algorithm to convert small sample data into large sample data through resampling, improving the prediction accuracy of the model and ensuring the reliability of the model. (3) This study evaluates the impact of each predictor variable on the performance of each model, which is more comprehensive and convincing than other machine models. (4) Based on the method of sample size verification, the relationship between sample size and AUC is explored, which provides a reference for sample analysis of predictive research.

Limitations of this study: (1) In terms of predictors, this study did not include laboratory and CT examination data, and its correlation needs to be further explored. (2) This study is a single-center study, and the research objects are limited to inpatients with COVID. There is a certain selection bias, and further multi-center and large-sample verification is required.

5 Conclusion

In conclusion, the integrated learning model has a good stopping effect on the risk of severe airflow limitation in patients with COVID. mMRC class, age, BMI, CAT score, smoking history and dyspnea are the main factors affecting airflow limitation. For those who are unable to perform lung functions, a severe airflow limitation risk warning model can help doctors assess patients' lung function and has great potential to effectively reduce future risks and burden for patients with COVID.

Table 7 Results of Risk Prediction Models for Airflow Limitation in Patients with COVID

Degree of airflow restriction Risk early warning model	model type	filling method	filter method	Number of variables	AUC	Accuracy	Precision	recall rate	F1Score
model 1	integrated learning	Not	Not	23	0.790 9	0.759 0	0.750 0	0.785 7	0.767 4
model 2	integrated learning	Not	Boruta	16	0.787 5	0.759 0	0.739 1	0.809 5	0.772 7
model 3	logistic regression	Not	Not	23	0.776 4	0.747 0	0.723 4	0.809 5	0.764 0
model 4	adaptive enhancement	Not	Lasso	4	0.773 8	0.698 8	0.680 9	0.761 9	0.719 1
model 5	integrated learning	Not	Lasso	4	0.773 8	0.698 8	0.680 9	0.761 9	0.719 1

Funding On Behalf of all authors the corresponding author states that they did not receive any funds for this project.

Data availability All the data is collected from the simulation reports of the software and tools used by the authors. Authors are working on implementing the same using real world data with appropriate permissions.

Declarations

Competing interests The authors declare that we have no competing interest.

References

1. Lei Ren, Dongqian Mi (2021) Prediction of COVID-19 Spread Based on Fractional SIR Model [J]. Adv Appl Mathematics 10(10):3233–3238. https://doi.org/10.12677/AAM.2021.1010338
2. Ruguo F, Yibo W, Ming L, Yingqing Z, Chaoping Z (2020) SEIR-based transmission model and inflection point prediction analysis of COVID-19 [J]. J Univ Electron Sci Technol China 49(3):369–374
3. Alenezi MN, Al-Anzi FS, Alabdulrazzaq H (2021) Building a sensible SIR estimation model for COVID-19 outspread in Kuwait. Alexan-dria Eng J 60:3161–3175. https://doi.org/10.1016/j.aej.2021.01.025
4. Chimmula VKR, Zhang L (2020) Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Soli-tons & Fractals 135:109864. https://doi.org/10.1016/j.chaos.2020.109864
5. Kırbaş İ, Sözen A, Tuncer AD, Kazancıoğlu FŞ (2020) Comparative analysis and forecasting of COVID-19 cases in various eu-ropean countries with ARIMA, NARNN and LSTM approaches. Chaos Solitons & Fractals 138:110015. https://doi.org/10.1016/j.chaos.2020.110015
6. Azarafza M, Azarafza M, Tanha J (2020) COVID-19 infection forecasting based on deep learning in Iran. MedRxiv. https://doi.org/10.1101/2020.05.16.20104182
7. Arora P, Kumar H, Panigrahi BK (2020) Prediction and analysis of COVID-19 positive cases using deep learning models: a descriptive case study of India. Chaos, Soli-tons & Fractals 139:110017. https://doi.org/10.1016/j.chaos.2020.110017
8. Omran NF, Abd-el Ghany SF, Saleh H, Ali AA, Gumaei A, Al-Rakhami M (2021) Applying deep learning methods on time-series data for forecasting COVID-19 in Egypt, Kuwait, and Saudi Arabia. Complexity 2021:6686745. https://doi.org/10.1155/2021/6686745
9. Verma H, Mandal S, Gupta A (2022) Temporal deep learning architecture for prediction of COVID-19 cases in India. Expert Systems App 195:116611. https://doi.org/10.1016/j.eswa.2022.116611
10. Zhang M, Chu R, Dong C, Wei J, Lu W, Xiong N (2021) Residual learning diagnosis detection: an advanced residual learning diagnosis detection system for COVID-19 in industrial internet of things. IEEE Trans Industr Inf 17(9):6510–6518. https://doi.org/10.1109/TII.2021.3051952
11. Chen Bin, Ai Chuan, Ma Liang et al (2020) Prediction of Epidemic T ediction of Epidemic Transmission and E ansmission and Evaluation of Pr aluation of Prevention and Contr ention and Control Measures Based on Artificial Society 32(12)
12. Ruguo F, Yibo W, Ming L et al (2020) SEIR-based covid-19 spread model and inflection point prediction analysis [J]. J Univ Electron Sci Technol China 49(3):369–374
13. Geng Hui Xu, Anding WX et al (2020) Analysis of the role of relevant intervention measures in the outbreak of novel coronavirus pneumonia based on SEIR model [J]. J Jinan Univ (Natural Sci Med) 41(2):175–180
14. Shengli C, Peihua F, Pengpeng S (2020) Modified SEIR infectious disease dynamics model applied to prediction and assessment of coronavirus disease 2019 (COVID-19) in Hubei Province [J]. J Zhejiang Univ (Med Sci) 49(2):178–184
15. Yang ZF, Zeng ZQ, Wang K et al (2020) Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions[J]. J Thorac Dis 12(3):165–174. https://doi.org/10.21037/jtd.2020.02.64
16. Zhixin W, Zhi L, Zhaojun L (2020) Analysis and prediction of novel coronavirus (COVID-19) epidemic based on machine learning [J]. Biomed Eng Res 39(1):1–5
17. Zhang L (2020) Fitness of the generalized growth to the COVID-19 data [J]. J Univ Electron Sci Technol China 49(3):345–348

18. Guo H, Yuping Z, Huanying H (2019) Application of seasonal ARIMA model in prediction of hand, foot and mouth disease epidemic in Jiangmen City [J]. China Health Stat 36(1):65–67
19. Hao Li, Deguang D, Xueqiang T et al (2020) Review of infectious disease dynamics model and its application in simulation and prediction of novel coronavirus pneumonia epidemic[J]. Med Health Equip 41(3):7–12
20. Lihong H, Yongyue W, Sipeng S et al (2020) Epidemic prediction methods and evaluation of common novel coronavirus pneumonia [J]. China Health Stat 37(3):322–326
21. Zhichao S (2012) Research on simulation technologies of the epidemics transmission and control based on artificial society[D]. National University of Defense Technology, Changsha
22. Junxiang T (2009) Research on AIDS transmission simulation modeling technology based on multi-agent and GIS integration[D]. Yunnan Normal Univ, Kunming
23. Kun Y, Jiangrong Li, Qingxiong C et al (2008) Research on the integrated model of AIDS spreading agent and GIS[J]. J Yunnan Normal Univ (Philosophy and Social Science Edition) 40(4):14–20