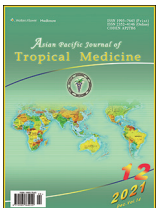

Original Article

Asian Pacific Journal of Tropical Medicine

apjtm.org

doi: 10.4103/1995-7645.332809

5-Year Impact Factor: 2.285

Predicting COVID-19 fatality rate based on age group using LSTM

Zahra Ramezani[1], Seyed Abbas Mousavi[2], Ghasem Oveis[3], Mohammad Reza Parsai[4], Fatemeh Abdollahi[5], Jamshid Yazdani Charati[1]✉

[1]*Department of Epidemiology and Biostatistics, School of Health, Mazandaran University of Medical Sciences, Sari, Iran*

[2]*Department of Psychiatry, Psychiatry and Behavioral Sciences Research Center, Addiction Institute, Mazandaran, University of Medical Sciences, Sari, Iran*

[3]*Health vice-chancellor of Mazandaran University of Medical Sciences, Sari, Iran*

[4]*Control Disease Center, Mazandaran University of Medical Sciences, Sari, Iran*

[5]*Department of Public Health, Psychiatry and Behavioral Sciences Research Center, Mazandaran University of Medical Sciences, Sari, Iran*

ABSTRACT

Objective: To predict the daily incidence and fatality rates based on long short-term memory (LSTM) in 4 age groups of COVID-19 patients in Mazandaran Province, Iran.

Methods: To predict the daily incidence and fatality rates by age groups, this epidemiological study was conducted based on the LSTM model. All data of COVID-19 disease were collected daily for training the LSTM model from February 22, 2020 to April 10, 2021 in the Mazandaran University of Medical Sciences. We defined 4 age groups, *i.e.*, patients under 29, between 30 and 49, between 50 and 59, and over 60 years old. Then, LSTM models were applied to predict the trend of daily incidence and fatality rates from 14 to 40 days in different age groups. The results of different methods were compared with each other.

Results: This study evaluated 50826 patients and 5109 deaths with COVID-19 daily in 20 cities of Mazandaran Province. Among the patients, 25240 were females (49.7%), and 25586 were males (50.3%). The predicted daily incidence rates on April 11, 2021 were 91.76, 155.84, 150.03, and 325.99 per 100000 people, respectively; for the fourteenth day April 24, 2021, the predicted daily incidence rates were 35.91, 92.90, 83.74, and 225.68 in each group per 100000 people. Furthermore, the predicted average daily incidence rates in 40 days for the 4 age groups were 34.25, 95.68, 76.43, and 210.80 per 100000 people, and the daily fatality rates were 8.38, 4.18, 3.40, 22.53 per 100000 people according to the established LSTM model. The findings demonstrated the daily incidence and fatality rates of 417.16 and 38.49 per 100000 people for all age groups over the next 40 days.

Conclusions: The results highlighted the proper performance of the LSTM model for predicting the daily incidence and fatality rates. It can clarify the path of spread or decline of the COVID-19 outbreak and the priority of vaccination in age groups.

KEYWORDS: COVID-19; Long short-term memory model; Incidence rate; Fatality rate; Prediction; Age classification

1. Introduction

The coronavirus disease caused by COVID-19 virus has urged many countries to control its spread through social distancing, masking, and determining the number of people who contact an infected person[1-3]. Many scientific and medical studies have investigated how to prevent its spread[4,5]. However, one of the most important issues is predicting the epidemic trend of

Significance

Previous studies have explored the risk factors and prediction models with the spread of SARS-CoV-2. What distinguishes this study from previous ones is age grouping to predict the COVID-19 trend. This study can clarify the path of spread or decrease in the daily incidence of the COVID-19 outbreak using LSTM and the priority of vaccination in different age groups.

✉To whom correspondence may be addressed. E-mail: jamshid.charati@gmail.com

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-Non Commercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: reprints@medknow.com

©2021 *Asian Pacific Journal of Tropical Medicine* Produced by Wolters Kluwer-Medknow. All rights reserved.

How to cite this article: Ramezani Z, Mousavi SA, Oveis G, Parsai MR, Abdollahi F, Charati JY. Predicting COVID-19 fatality rate based on age group using LSTM. Asian Pac J Trop Med 2021; 14(12): 564-574.

Article history: Received 4 July 2021 Revision 6 December 2021
Accepted 9 December 2021 Available online 29 December 2021

COVID-19[6,7]. Although traditional time series methods work well in time-dependent sequence observations, they have many limitations. For example, outliers can cause biased estimation of model parameters; when a large number is estimated, direct human intervention and evaluation are necessary to select the final model[8]. Time series models are often linear; they might not be able to explain nonlinear behavior well. Many traditional statistical methods do not learn new data entry well; they require periodical reevaluation. Neural networks can overcome these limitations, or at least they have fewer problems compared to traditional time series statistical methods[9]. Although they are inherently nonlinear, they are also able to model linear patterns[8,10].

Kırbaş *et al.* performed a comparative analysis in Turkey and employed AutoRegressive Integrated Moving Average (ARIMA), Nonlinear AutoRegressive Neural Network (NARNN), and long short-term memory (LSTM) methods to model the COVID-19 confirmed cases in Denmark, Belgium, Germany, France, the United Kingdom, Finland, Switzerland, and Turkey. They used six model performance metrics (*i.e.*, MSE, PSNR, NMSE, MAPE, and SMAPE) to choose the most precise model. The results of the first stage of their study confirmed LSTM as the most precise model. However, the second stage revealed that it was successful in predicting a 14-day view. It showed that the growth rate would slightly drop in many countries[11]. In 2020, Arora *et al*. conducted a study to predict and analyze positive cases of COVID-19 using deep learning-based models in India. To achieve their goal, they employed different LSTM models based on recurrent neural networks (RNNs), including Deep LSTM, Convolutional LSTM, and Bi-directional LSTM. Finally, they selected the LSTM model with minimal error to predict the daily and weekly cases[12]. Moreover, Rashed *et al*. proposed an LSTM architecture to predict the spread of COVID-19 considering various factors such as public mobility estimates and meteorological data; finally, they applied it to the data collected in Japan. They predicted the positive cases in six prefectures of Japan for different time frames[13]. Other studies have been performed for forecasting new cases and deaths consisting of vanilla, stacked, bidirectional, and multilayer LSTM models. Chatterjee *et al*. tried to limit the exponential spread to slow down the transmission rate (spread factor) and then assessed the risk factors associated with COVID-19. However, the results indicate that vanilla, bidirectional, and stacked LSTM models outperformed multilayer LSTM models[14]. Albahli *et al*. applied a semantic analysis of three levels (negative, neutral, and positive) to measure the people's feelings towards the pandemic and lockdown in the Gulf countries[15]. In another study by Odhiambo *et al*., an RNN within LSTM was compared to the traditional ARIMA method in countries with limited data availability, such as Kenya. The results demonstrated that the LSTM network was precise when forecasting the future systematic fatality risks compared to the traditional time series method[16].

Unlike previous studies, we predict the daily incidence and fatality rates in each age group in detail. The daily incidence rate is the proportion of the number of cases to the total population multiplied by WHO Standard Population per 100 000 people. Also, the fatality rate is the proportion of the number of fatality to the total population multiplied by WHO Standard Population per 100 000 people. The advantage of this study is predicting the daily incidence and fatality rates of COVID-19 cases in different age groups based on different populations by LSTM in areas near the Caspian Sea. In this way, a proper decision can be made to prevent the spread of the disease and prioritize vaccination. To predict the daily incidence and fatality rates from 14 to 40 days for each age group, we focused our analysis on the data recorded by Mazandaran University of Medical Sciences.

2. Materials and methods

2.1. Study design and data collection

To predict the daily incidence and fatality rates of COVID-19 by age groups in Mazandaran Province, diagrams and descriptive statistics tables have been used to describe the existing conditions. This could help us to investigate the effect of age on the increased daily incidence and fatality rate. Then, the groups have been compared in terms of prevalence and prediction of daily incidence and fatality rate. As for modeling, we attempted to predict the daily incidence and fatality rate daily and monthly. Thus, the data have been collected for training based on 50 826 admitted patients and 5 109 deaths of COVID-19 in 20 cities in Mazandaran Province from February 22, 2020 to April 10, 2021. After we prepared the data, regression coefficients, confidence interval, correlation heatmap, and comparison graphs for daily incidence and fatality rate were presented for clear descriptions and better decision making. Then, the traditional ARIMA model and the LSTM models have been implemented for forecasting.

2.2. Proposed model

We used an expert-based standard checklist to collect data, including disease symptoms, demographic characteristics, history of disease, and other risk factors. This study attempted to predict the daily incidence and fatality rates in Mazandaran Province based on WHO standard population[17].

Due to the time series data and the large volume of data, we could use the LSTM networks, widely applicable in time-dependent

studies, for forecasting. Statistical analyses were done by SPSS software version 26 and Python software version 3.7.

The LSTM model is an RNN in which the prediction result for the next time unit is based on the current situation and previous knowledge[18,19]. This can also consider short-term and long-term correlations within the time series in the LSTM network by using the hidden layer as a memory block, which can learn long-term dependencies of the content[20]. Each LSTM cell consists of input, output, and forget gates in a hidden layer. The LSTM cell internal memory stores only useful and relevant information. Figure 1 depicts the structure of an LSTM network with 3 gates. The LSTM network is defined using the following equations:

$$f_t = \sigma\,(W_f . [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma\,(W_i . [h_{t-1}, x_t] + b_i)$$
$$\tilde{c}_t = tanh\,(W_c . [h_{t-1}, x_t] + b_c)$$
$$o_t = \sigma\,(W_o . [h_{t-1}, x_t] + b_o)$$
$$c_t = f_t . c_{t-1} + i_t . \tilde{c}_t$$
$$h_t = o_t * tanh\,(c_t)$$

Where x_t and h_t are input and output vectors, respectively, f_t is a forget gate vector, c_t represents the cell state vector, i_t is the input gate vector. o_t is the output gate vector, and W and b show the parameter matrices.

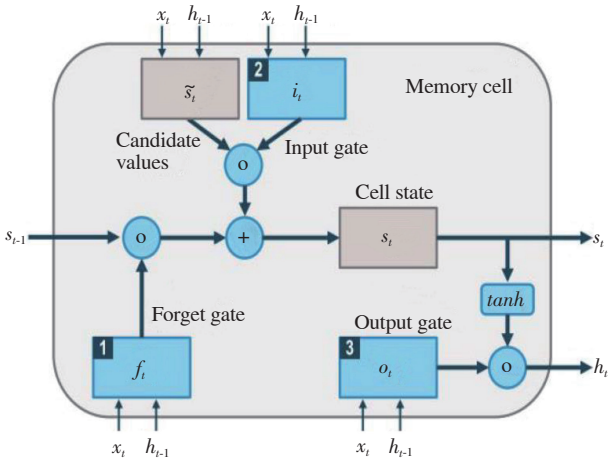

Figure 1. The presentation of LSTM memory cell structure follows Fischer and Krauss[10].

By assigning different functions to gates, the LSTM memory block can record complex features correlations in short-term and long-term time series; it is a significant advantage over RNN[21]. We should note that other appropriate transformations may be used if necessary to establish conditions and assumptions along with better estimates. The data are divided into two datasets of training

Table 1. Clinical characteristics and outcomes of patients referred for COVID-19 treatment in 20 cities of Mazandaran Province from February 22, 2020 to April 10, 2021.

Characteristics	Total (n= 50826)	Alive (n= 45717)	Dead (n= 5109)	P-value
Gender, n (%)				0.126
Male	25586 (50.3)	23240 (50.8)	2346 (45.9)	
Age category, years, n (%)				<0.001
≤29	6215 (12.2)	6074 (13.3)	141 (2.8)	
30-49	13832 (27.2)	13342 (29.2)	490 (9.6)	
50-59	8401 (16.5)	7725 (16.9)	676 (13.2)	
≥60	22378 (44.0)	18576 (40.6)	3802 (74.4)	
ICU				<0.001
Yes	6953 (13.7)	5116 (11.2)	1837 (36.0)	
Lung disease				<0.001
Yes	2060 (4.1)	1734 (3.8)	326 (6.4)	
Chronic renal disease				<0.001
Yes	1990 (3.9)	1608 (3.5)	382 (7.5)	
Cardiovascular disease				<0.001
Yes	3595 (7.1)	2916 (6.4)	679 (13.3)	
Other diseases				<0.001
Yes	6902 (13.6)	5986 (13.1)	916 (17.9)	
Chronic liver disease				0.390
Yes	139 (0.3)	105 (0.2)	34 (0.4)	
Diabetes				<0.001
Yes	9374 (18.4)	7982 (17.5)	1392 (27.2)	
Cough				<0.001
Yes	21495 (42.3)	19507 (42.7)	1988 (38.9)	
Shortness of breath				0.667
Yes	25462 (50.1)	22113 (48.4)	3349 (65.6)	
Sore throat				<0.001
Yes	6307 (12.4)	5816 (12.7)	491 (9.6)	
Headache				<0.001
Yes	5034 (9.9)	4704 (10.3)	330 (6.5)	
Diarrhea				<0.001
Yes	2353 (4.6)	2225 (4.9)	128 (2.5)	

and testing, and finally, the prediction occurs through experimental data. The purpose of normalization is generally to reduce the computation time due to the shrinking of the numbers. The mean squared logarithmic error (MSLE) criteria and Adam optimizer are chosen for better forecasting. The lower the value, the better the model estimate.

3. Results

Before predicting the daily incidence and fatality rates, we compared different age groups according to the available data from 20 cities in Mazandaran Province. COVID-19 case data were recorded daily from February 22, 2020 to April 10, 2021 in Mazandaran University of Medical Sciences. The daily incidence and fatality rates of different age groups were calculated daily according to the WHO World Standard Population.

Table 1 indicates the characteristics and behavior of COVID-19. Among the patients infected with this virus, 25 240 were females (49.7%), and 25 586 were males (50.3%). A total of 5 109 patients died, among which 2 763 (54.1%) were women and 2 346 (45.9%) were men. Table 2 shows the population of the province in age groups and the population of urban/rural men and women. Of the total population, 1 581 594 were urban and 1 175 263 were rural. We classified the data into 4 age groups, patients under 29, between 30 and 49, between 50 and 59, and over 60 years old in Table 2. The *P*-value is calculated based on the *Chi*-square test among the 4 age groups ($P<0.001$).

Table 2. Population of 20 cities of Mazandaran Province based on demographic characteristics.

Characteristics	*n* (%)
Age, years	
≤29	1 272 531 (46.2)
30-49	913 960 (33.1)
50-59	284 314 (10.3)
≥60	286 052 (10.4)
Rural population	
Female	578 086 (49.2)
Male	597 177 (50.8)
Urban population	
Female	776 960 (49.1)
Male	804 634 (50.9)

In the following, we analyzed the collected data to identify patterns and trends. Table 3 examines the effects of several specific disease histories on the fatality age of COVID-19 patients. The coefficient estimates the marginal effect of a one-unit increase (a disease) in that independent variable on the dependent variable (age category), holding constant all other variables in the model. According to the disease history of the people of the region, the results show that the effects of cardiovascular and diabetic diseases and other diseases, including asthma, have the greatest impact on the age categories. It is shown that COVID-19 patients with diabetes (the regression coefficient 0.545) were at a higher risk in the age groups. Although the coefficient in the model on cardiovascular disease (the regression coefficient 0.610) is larger than the coefficient on diabetes (the regression coefficient 0.545), it does not make sense to compare those coefficients directly. Other diseases, including asthma, are in the next ranks in terms of regression coefficients. Also, smaller regression coefficients have a lesser effect on the age categories. The negative coefficient of liver disease is due to the low frequency of this disease in the study population or the lack of registration of this type of disease in COVID-19 patients. Regression coefficients and confidence intervals were presented for considering significance level in Table 3. The history of various diseases is significant for $P<0.001$, such as diabetes.

The correlation heatmap of real COVID-19 data is depicted in Figure 2. As age increases, the number of new fatalities increases due to the high correlation value of |r|.

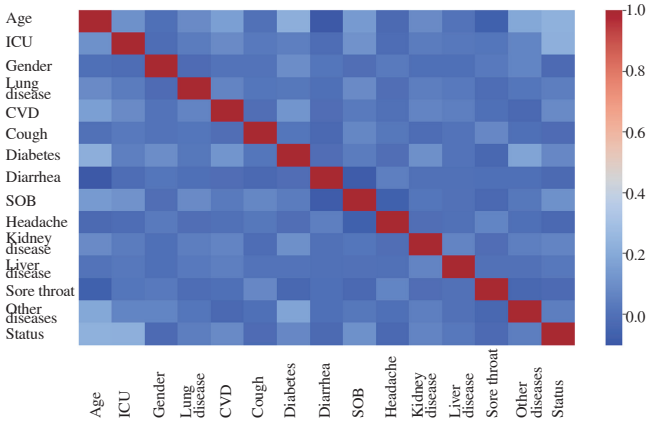

Figure 2. Correlation heatmap of real data. ICU: intensive care unit; CVD: cardiovascular disease; SOB: shortness of breath.

Table 3. The effect of comorbidity of COVID-19 patients on the dependent variable of age categories using regression coefficients and 95% confidence interval.

Model	Unstandardized coefficients		Standardized β	*t*	*P*	95% *CI* for β	
	β	Std. error				Lower bound	Upper bound
(Constant)	2.735	0.007		392.192	<0.001	2.721	2.749
Diabetes	0.545	0.012	0.193	44.739	<0.001	0.521	0.569
kidney disease	0.312	0.024	0.055	13.099	<0.001	0.265	0.359
Liver disease	-0.076	0.088	-0.004	-0.860	0.390	-0.248	0.097
Other diseases	0.531	0.014	0.166	38.951	<0.001	0.504	0.558
Cardiovascular disease	0.610	0.018	0.143	33.756	<0.001	0.574	0.645

Age categories: ≤29 years, 30-49 years, 50-59 years, ≥60 years. 95% *CI*: 95% confidence interval.

Figure 3 shows the average daily incidence and fatality rates of the groups based on the World Standard Population per 100 000 people. Figure 4A shows the daily incidence trend of the registered cases in 20 cities in Mazandaran Province. Figure 4B and 4C show the evaluation of the daily incidence and fatality rate for each age groups in Mazandaran Province regarding the population per 100 000 people. As shown in Figure 4C which evaluates and compares the COVID-19 fatality rate in 4 age groups in Mazandaran Province, patients over 60 and between 50 and 59 have the highest fatality rate according to the WHO World Standard Population.

3.1. Time series ARIMA model

ARIMA is a time series prediction model which is a form of regression analysis and is used to forecast the future trends in the time series dataset. This model is applied to capture the autocorrelation from the data which computes the future values based on the correlations between the previous values. A traditional ARIMA model has been implemented to the COVID-19 data before considering the LSTM model. Then, the predicted results of COVID-19 cases using the ARIMA model have been presented.

At first, the Dickey-Fuller test is used to examine if the time series is stationary. The null hypothesis (H0) was rejected with a *P*-value ≤ 0.05 in the Dickey-Fuller test, indicating that the data do not have a unit root and are stationary. If the test statistic is less than critical values, we reject the null hypothesis. If the test statistic is greater than critical values, we accept the null hypothesis.

Here, the test statistics value=-2.83 is greater than the critical value(1%)=-3.45 and the critical value(5%)=-2.87, thus the data is not stationary. The test statistic is less than the critical value(10%)=-2.57 and the data is stationary.

We have to transform the data to make the data more stationary for critical value 1% and critical value 5%. But, the data are stationary in significant value 10% and we apply the ARIMA model for a significant value 10%. An ARIMA statistical model has been used to predict the daily incidence trend of the COVID-19 outbreak in the time series (Figure 5).

Note that for a series to be stationary, it must follow some principles such as modeling, estimating trends, and seasonal changes in the series, along with their removal from the series. Then, the forecasting techniques can be implemented in the data. In the following, it can be seen that the LSTM models do not have the complexities of traditional time series methods and produce more accurate results and are closer to the actual data.

3.2. LSTM model

We have illustrated applied hyper-parameters, various LSTM models, and loss functions to consider the proposed model in this section.

Optimizer explores specific configurations to speed up or slow down learning that leads to benefits. Adam optimizer applies the learning rate of 0.001, provides a reliable method in the stochastic gradient descent algorithm, and computes adaptive learning rates for each parameter. The 50 epochs have been specified for observing the loss curve during training and convergence of the loss curve. The main hyper-parameters, including the sequence length, activation function, learning rate, batch size, epochs, optimizer, loss function, and n_hidden, are listed in Table 4.

Table 4. Hyper-parameter used in the LSTM model.

Hyper-parameter	Value
Sequence length	50
Activation function	ReLU
Learning rate	0.001
Batch_size	32
Epochs	50
Optimizer	Adam
Loss function	MSLE
n_hidden	32

The training set is 85% of the data, while the remaining 15% are applied as testing set[11]. We considered an approximately 14- to 40-day prediction period for testing data. More specifically, the data were split into two subsets. The first subset was composed of training (from February 22, 2020 to April 10, 2021) and test

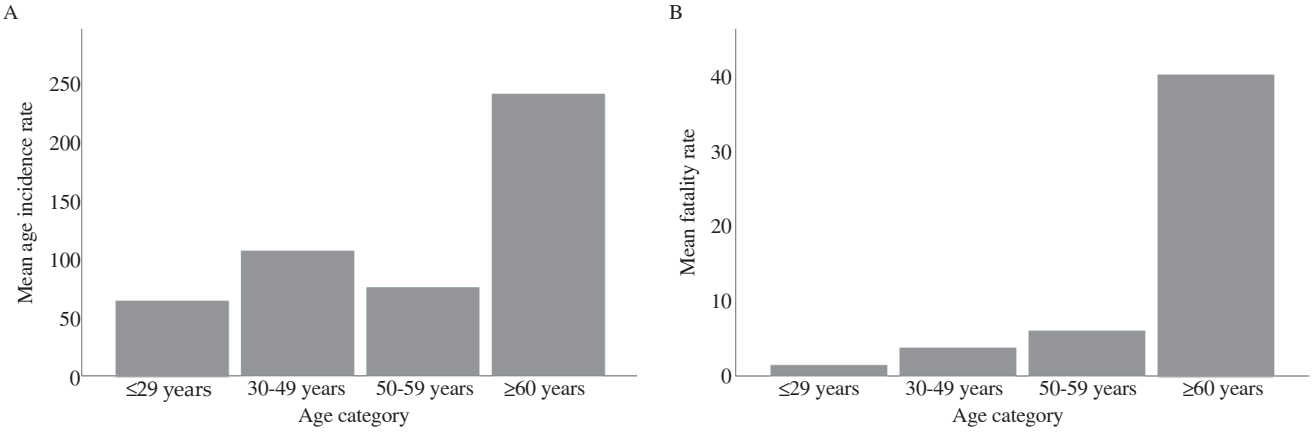

Figure 3. Mean comparison of COVID-19 incidence rate and fatality rate in 4 age groups according to the WHO World Standard Population in 20 cities of Mazandaran Province from February 22, 2020 to April 10, 2021. Simple bar mean of age (A) incidence rate and (B) fatality rate by category.

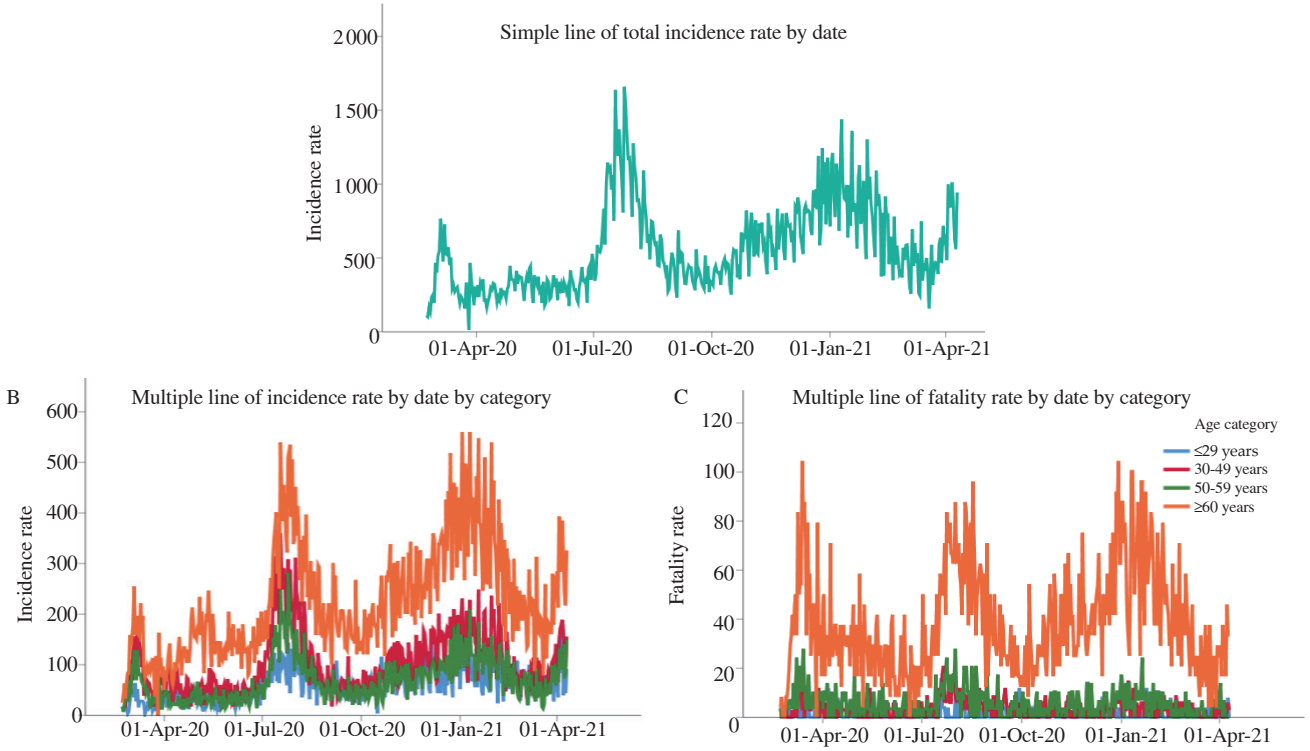

Figure 4. The trend of COVID-19 outbreaks. (A) The general incidence rate of COVID-19 patients for 20 cities in Mazandaran Province; (B)Comparison of COVID-19 incidence rate in 4 age groups; (C) Comparison of COVID-19 patients' fatality rate in 4 age groups.

data (the last 14 days, from April 11, 2021 to April 24, 2021). On the other hand, the second subset was composed of training (from February 22, 2020 to April 10, 2021) and test data (the last 40 days, from April 11, 2021 to May 25, 2021) for prediction analysis.

Table 5 illustrates the average performance results of various LSTM models. In this study, the differences in various loss values between models are insignificant due to the sufficient data availability and a more detailed investigation in each age group. Although the results show that vanilla, stacked, and bidirectional LSTM models outperform other LSTM models, we selected a simple LSTM model for faster training and prediction with lower loss. An MSLE loss function was selected as the suitable metric to train to predict the daily incidence and fatality rates in the LSTM model. For models without data grouping, selecting stacked LSTM is more appropriate due to being a deeper model.

Table 5. Comparative analysis of various LSTM models in terms of error metrics for predicting COVID-19 outbreak in Mazandaran Province.

LSTM model(s)	MAE	MSE	MSLE	R^2 score
Vanilla	0.3587	0.5155	0.1622	0.98
Stacked	0.3984	0.5700	0.1644	0.97
Bidirectional	0.3271	0.5202	0.1132	0.98
Multi-layer 1	0.6818	0.9696	0.0774	0.91
Multi-layer 2	0.7545	0.9781	0.0536	0.91

Daily incidence and fatality rates of real data have been evaluated in Table 6 from March 20, 2020 (March 20 is the first day of the first month of the year in Iran) for 12 consecutive months. Since, in the first days of the disease outbreak in the country, the data were not well recorded or the disease was not diagnosed, the daily incidence and fatality rate of real data have been calculated from

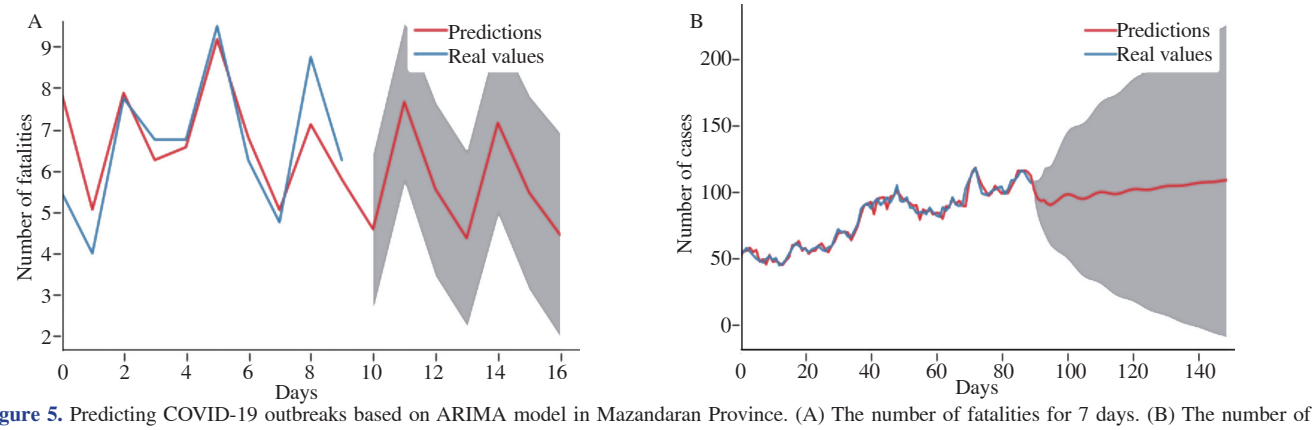

Figure 5. Predicting COVID-19 outbreaks based on ARIMA model in Mazandaran Province. (A) The number of fatalities for 7 days. (B) The number of confirmed cases for 40 days. The shade is the prediction interval that predicts in what range a future observation will put.

March 20, 2020. In general, training data from February 22, 2020 (*i.e.*, the first recorded data) have been used daily to predict the COVID-19 outbreaks using the LSTM model.

Table 6 separately displays the COVID-19 daily incidence rate for 12 consecutive months in 4 groups. For example, the 10th month has the highest daily incidence rate, and the vulnerable class of the category of over 60 years old has the highest rate of 405.53 person per 100 000 people. In the same way, Table 6 also depicts the fatality rate in each age group for 12 consecutive months, indicating a trend similar to the daily incidence rate in the groups.

Training data from February 22, 2020 to April 10, 2021 were trained by LSTM architecture. Figure 6 shows the trend of loss function values of training and validation to predict the confirmed cases and fatality rate in the two age groups as examples. Moreover, similar results have been achieved for other groups. Predictions of group 1 are related to under 29, group 2 between 30 and 49, group 3 between 50 and 59, and group 4 over 60 years old. Then, we predicted the daily incidence and fatality rates for 14 to 40 days from April 11, 2021.

Figure 7 and 8 illustrate the performance of the proposed model and prediction by age groups in Mazandaran Province. Figure 7 shows the predicted values of the COVID-19 patients in Mazandaran Province by 4 age groups with the LSTM model for 14 days after the last date of the training. On the other hand, Figure 8 depicts the prediction of the daily incidence rate of Mazandaran Province for 4 age groups in 40 days by the LSTM model. Before the vertical line, the trend of the training data daily incidence rate before April 11, 2021 is shown, and the trend of predicting the daily incidence rate can be observed after this line.

Table 7 shows the prediction of cases and daily incidence

Table 6. Daily incidence rate of confirmed cases and fatality rate of real data in each age group in per 100 000 people for 12 consecutive months.

Variables	Month ID	Date	Age categories				Total
			≤29 years	30-49 years	50-59 years	≥60 years	
Incidence rate	1	March 20, 2020-April 19, 2020	31.27	57.13	28.70	70.69	187.79
	2	April 20, 2020-May 20, 2020	28.31	53.56	34.67	158.95	275.49
	3	May 21, 2020-June 20, 2020	31.79	43.89	31.52	133.87	241.07
	4	June 21, 2020-July 21, 2020	69.24	153.81	97.93	244.69	565.67
	5	July 22, 2020-August 21, 2020	79.54	174.20	123.71	335.02	712.47
	6	August 22, 2020-September 21, 2020	54.31	64.96	49.64	189.96	358.87
	7	September 22, 2020-October 21, 2020	52.40	64.63	44.78	175.53	337.34
	8	October 22, 2020-November 20, 2020	73.28	101.69	72.11	244.77	491.85
	9	November 21, 2020-December 20, 2020	84.58	133.06	88.63	292.55	598.82
	10	December 21, 2020-January 19, 2021	95.49	160.83	121.32	405.53	783.17
	11	January 20, 2021-February 18, 2021	78.86	130.96	103.98	320.41	634.21
	12	February 19, 2021-March 20, 2021	61.18	70.83	52.92	193.92	378.85
Fatality rate	1	March 20, 2020-April 19, 2020	2.32	4.35	6.08	33.97	46.72
	2	April 20, 2020-May 20, 2020	1.03	2.42	4.84	29.39	37.68
	3	May 21, 2020-June 20, 2020	0.64	2.03	4.39	21.17	28.23
	4	June 21, 2020-July 21, 2020	1.16	3.96	4.61	27.64	37.37
	5	July 22, 2020-August 21, 2020	1.93	7.93	12.61	65.66	88.13
	6	August 22, 2020-September 21, 2020	1.29	4.93	4.73	36.94	47.89
	7	September 22, 2020-October 21, 2020	2.39	2.30	3.84	22.85	31.38
	8	October 22, 2020-November 20, 2020	1.20	2.50	4.77	37.75	46.22
	9	November 21, 2020-December 20, 2020	1.46	2.60	5.00	42.49	51.55
	10	December 21, 2020-January 19, 2021	1.20	4.89	8.14	66.87	81.10
	11	January 20, 2021-February 18, 2021	1.73	3.30	6.28	55.58	66.89
	12	February 19, 2021-March 20, 2021	0.53	2.70	2.33	31.07	36.63

Table 7. Predicting COVID-19 cases and daily incidence rate for the 4 age groups in 14 consecutive days from April 11, 2021 to April 24, 2021 as per 100 000 people.

Forecast Id	Date	No. of cases				Incidence rate			
		≤29 years	30-49 years	50-59 years	≥60 years	≤29 years	30-49 years	50-59 years	≥60 years
1	April 11, 2021	23	52	43	78	91.76	155.84	150.03	325.99
2	April 12, 2021	17	48	34	69	67.82	143.85	118.63	288.37
3	April 13, 2021	23	50	29	70	91.76	149.84	101.18	292.56
4	April 14, 2021	19	47	27	67	75.80	140.86	94.20	280.01
5	April 15, 2021	13	45	30	65	51.86	134.86	104.67	271.66
6	April 16, 2021	16	39	25	57	63.83	116.88	87.23	238.22
7	April 17, 2021	12	32	26	57	47.88	95.90	90.72	238.22
8	April 18, 2021	16	42	31	67	63.83	125.87	108.16	280.01
9	April 19, 2021	13	40	29	60	51.86	119.87	101.18	250.76
10	April 20, 2021	14	41	26	60	55.86	122.87	90.72	250.76
11	April 21, 2021	12	38	25	57	47.87	113.88	87.23	238.22
12	April 22, 2021	10	38	25	56	39.90	113.88	87.23	234.04
13	April 23, 2021	11	35	23	54	43.89	104.89	80.25	225.68
14	April 24, 2021	9	31	24	54	35.91	92.90	83.74	225.68

rates for the four groups from April 11, 2021 to April 24, 2021 for 14 consecutive days. For a simpler and more meaningful representation of the prediction values for 40 consecutive days, we have shown the prediction trend of daily COVID-19 cases in Figure 8 for all 4 groups.

In addition, the average predicted values of daily incidence and fatality rates for 40 days have been shown for each age category in Table 8. Predictions in stable conditions are very close to the actual values[22].

Table 8. Predicting average daily incidence and fatality rate of COVID-19 outbreak for 40 days in each age category as per 100 000 people in Mazandaran Province.

Variable	Category				Total
	≤29	30-49	50-59	≥60	
Incidence rate	34.25	95.68	76.43	210.80	417.16
Fatalities rate	8.38	4.18	3.40	22.53	38.49

The incidence and fatality rate were calculated as per 100 000 people.

4. Discussion

Previous studies have mainly focused on the effective factors such as age, underlying diseases, and fatality rate of COVID-19[23,24]. Moreover, they investigated the COVID-19 disease predictions and fatality rate regardless of the incidence rate in age groups[12,25]. For example, Sasson showed that the age pattern of COVID-19 fatality in different countries might indicate a difference in population health, clinical care standards, or data quality[26].

Researchers have shown that COVID-19 is very common in elderly patients with underlying diseases such as cardiovascular disease, high blood pressure, and diabetes. Due to the diversity in the demographic statistics, underlying diseases, and health systems, the fatality rate of COVID-19 disease was predicted for 187 countries, ranging from 0.43% in Sub-Saharan Africa to 1.45% in Eastern Europe[27].

What distinguishes this research from other studies is the accurate prediction of incidence and fatality rates by different age groups

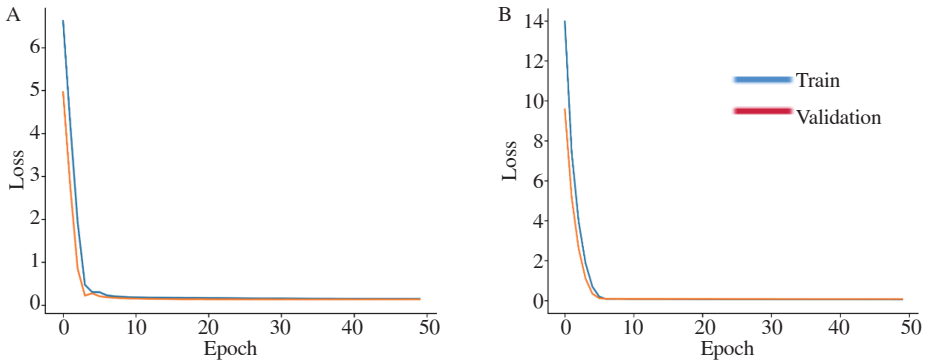

Figure 6. Loss function diagram of LSTM model to predict the number of confirmed cases and fatality rate in age category of (A) ≤29 years and (B) above 60 years as examples.

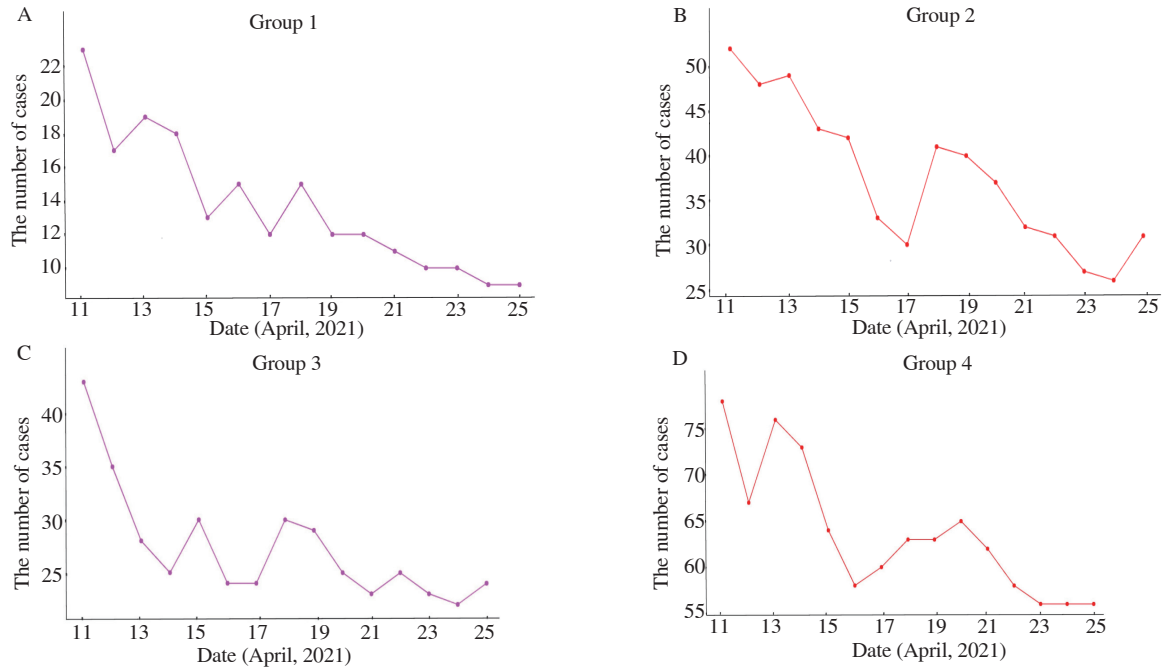

Figure 7. Prediction of the COVID-19 cases in Mazandaran Province by 4 age groups with the LSTM model in 14 days from April 11, 2021 to April 24, 2021. Group 1: ≤29 years; Group 2: 30-49 years; Group 3: 50-59 years; Group 4: ≥60 years.

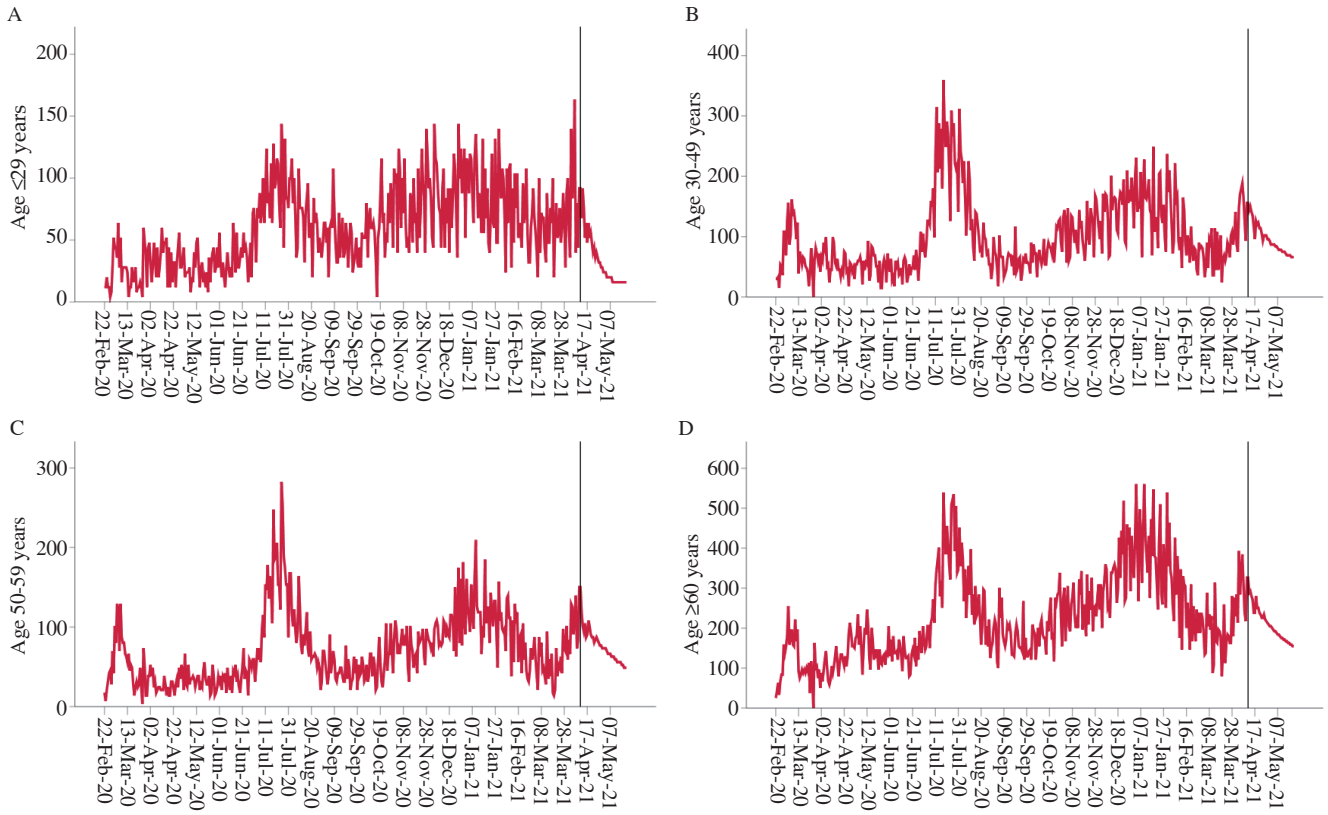

Figure 8. Predicting the trend of incidence rate COVID-19 outbreak for 40 days in 4 age groups by the LSTM model in Mazandaran Province. (A) Incidence rate of age ≤29 (Y-axis) *vs.* days (X-axis); (B) Incidence rate of age between 30-49 (Y-axis) *vs.* days (X-axis); (C)Incidence rate of age between 50-59 (Y-axis) *vs.* days (X-axis). (D) Incidence rate of age ≥ 60 (Y-axis) *vs.* days (X-axis). The vertical line separates the incidence rate trend of the previous days and the prediction trend of incidence rate.

using the LSTM deep learning technology. Furthermore, we achieved accurate results compared to those who worked on the general case disregarding age grouping. Thus, the diagnosis of the high-risk age group and the predicted values illuminates the future of the disease outbreak.

A meta-analysis with a large number of patients highlights the determining effect of age on fatality. The data of this study were collected from the patients in 20 cities near the Caspian Sea in Mazandaran Province, and the daily incidence and fatality rates of each age group were predicted in detail. Due to the time series data and their large volume, the researchers selected LSTM networks, widely applicable in the study of time-dependent issues for forecasting.

Evaluation metrics are loss functions such as mean absolute error (MAE), mean squared error (MSE), mean squared logarithmic error (MSLE), binary cross-entropy, categorical cross-entropy, residual forecast error/forecast error, forecast bias/mean forecast error, root mean square error (RMSE), and R^2 score as adjusted *R*-squared for the model. To assess individual regression models, we applied MAE, MSE, MSLE, and R^2 regression metrics. The LSTM model is compiled with Adam optimizer, loss function of MSLE, and accuracy.

In a comparative study with national reports data on May 7, 2020, from China, Italy, Spain, the United Kingdom, and New York State, Bonanad *et al*. showed an overall fatality rate of 12.10%. The fatality rate changes between countries with the relevant thresholds on age >50 and age >60 years old. The lowest fatality rate was in China (3.1%), and the highest was in the United Kingdom (20.8%) and New York State (20.99%). The fatality rate was <1.1% in patients aged <50 years, and it has exponentially increased in older ones in the recorded data in 5 countries. Besides, the highest fatality rate occurred in patients aged 80 years[24].

This study scrutinized 50 826 COVID-19 patients with 5 109 deaths in 20 cities of Mazandaran Province from February 22, 2020 to April 10, 2021. The researchers assessed the mean standardized incidence and fatality rates by age group based on training data available for 12 months. The results revealed that in each age group, that is, patients under 29, between 30 and 49, between 50 and 59, and over 60 years old, the standard incidence rates per 100 000 people were 31.27, 57.13, 28.70, and 70.69 in the first month, respectively. In the 12th month, the standard incidence rates were 61.18, 70.83, 52.92, and 193.92 in each age group, respectively. Moreover, the fatality rates in each age group in the first month were 2.32, 4.35, 6.08, and 33.97 per 100 000 people, while in the 11th month it was 1.73, 3.30, 6.28, and 55.58, and in the 12th month, it was 0.53, 2.70, 2.33, and 31.07 per 100 000 people.

The results demonstrate the daily incidence rates fluctuations

in different months and the increase in the incidence rates with the increase in age. In addition, we obtained the daily number of incidence and fatality by age groups. Finally, we predicted the standard incidence and fatality rates in each age group for the next 14 to 40 days. The prediction values were close to the real values. The daily incidence rates in April 11, 2021 were at 91.76, 155.84, 150.03, and 325.99 per 100 000 people, respectively. In general, the average standard daily incidence rate for 4 age groups per 100 000 people were 34.25, 95.68, 76.43, and 210.80 for the next 40 days, respectively. Correspondingly, the average daily fatality rate for the 4 age groups were 8.38, 4.18, 3.40, and 22.53, respectively.

Although a fixed parameter cannot be a single factor, COVID-19 infections are inherently associated with the age pattern. In this article, all indices were based on the WHO standard population. We also underestimated our calculations; that is, the patients with mild COVID-19 had not been included in the study. Overall, the results show that COVID-19 is life-threatening not only for older adults but for middle-aged people, and the high or low risk is predictable in the coming days.

Similar to any other study, this research is subject to several limitations. First, model training with more data leads to better results when compared to different countries. In addition, the accuracy of the LSTM prediction improves after considering more parameters instead of relying on the univariate trend of time series data. Currently, this model can predict 14 to 40 days with acceptable accuracy. Moreover, we had an underestimation in the calculation due to not including the mild disease in the study. According to the purpose of the study, *i.e.*, predicting the growth of coronavirus disease in different age groups, we applied the LSTM models. Since the results were obtained with limited data availability (*i.e.*, 20 cities near the Caspian Sea in Mazandaran province), the researchers used the results of the other studies conducted in different countries. However, information on transmission distance based on different variants was not available due to the lack of appropriate technology. This can be a recommended issue to be studied in the future, considering different age groups.

The results show that the main priority in the preventive measures should be older patients who are more susceptible to this disease. If public health proceedings reduce infection in the old patients, it can significantly reduce fatality. By predicting the number of admitted patients and the fatality and incidence rates of patients in each age group, we can prevent COVID-19 prevalence.

In conclusion, we predicted COVID-19 incidence and fatality rates by age groups using the LSTM network based on the WHO population. The LSTM network predicted the number of confirmed cases and incidence and fatality rates in 14 to 40 days. For example, the incidence rate for over 60 years old patients was obtained 210.80 per 100 000 people. The results showed that the incidence and fatality rates of COVID-19 patients in Mazandaran Province in the age group of 60 years and above are higher than other groups. The prediction results show fluctuations in the incidence and fatality rates, though the values are accurately predicted for each age group. By differentiating age groups in predicting the number or rates of incidence and fatality, the researchers obtained accurate results compared to predictions without differentiating groups. Predicting the incidence and fatality rates of different groups, we can make better decisions about the essential health proceedings as well as vaccination prioritization.

Conflict of interest statement

The authors declare that there is no conflict of interest.

Authors' contributions

ZR performed research, designed the analysis, implemented python programming, analyzed, interpreted the data, wrote and revised the manuscript. SAM contributed to COVID-19 data acquisition. GO participated in the discussion. MRP contributed to COVID-19 data acquisition. FA participated in the discussion. JYCh designed research, contributed to the interpretation and edited the manuscript.

References

[1] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020; **579**(7798): 265-269.

[2] Nguyen TT, Nguyen QVH, Nguyen DT, Hsu EB, Yang S, Eklund P. Artificial intelligence in the battle against coronavirus (COVID-19): A survey and future research directions. *arXiv: 2008.07343* 2020. doi: 10.13140/RG.2.2.36491.23846/1.

[3] Anjorin AA. The coronavirus disease 2019 (COVID-19) pandemic: A review and an update on cases in Africa. *Asian Pac J Trop Med* 2020; **13**(5): 199.

[4] Jahanbin K, Rahmanian V. Using twitter and web news mining to predict COVID-19 outbreak. *Asian Pac J Trop Med* 2020; **13**(8): 378.

[5] Islam MZ, Islam MM, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics Med Unlocked* 2020; **20**: 100412.

[6] Pereira IG, Guerin JM, Junior AGS, Distante C, Garcia GS, Goncalves LMG. Forecasting Covid-19 dynamics in Brazil: A data driven

approach. *arXiv: 2005.09475* 2020. doi: 10.3390/ijerph17145115.

[7] Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, 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.

[8] Zhang GP, Qi M. Neural network forecasting for seasonal and trend time series. *Eur J Oper Res* 2005; **160**(2): 501-514.

[9] Hill T, O'Connor M, Remus W. Neural network models for time series forecasts. *Manage Sci* 1996; **42**(7): 1082-1092.

[10]Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. *Eur J Oper Res* 2018; **270**(2): 654-669.

[11]Kırbaş İ, Sözen A, Tuncer AD, Kazancıo FŞ. Comperative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos, Solitons & Fractals* 2020; **138**: 110015. doi: 10.1016/j.chaos.2020.110015.

[12]Arora P, Kumar H, Panigrahi BK. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals* 2020; **139**: 110017.

[13]Rashed EA, Hirata A. One-year lesson: Machine learning prediction of COVID-19 positive cases with meteorological data and mobility estimate in Japan. *Int J Environ Res Public Health* 2021; **18**(11): 5736.

[14]Chatterjee A, Gerdes MW, Martinez SG. Statistical explorations and univariate timeseries analysis on COVID-19 datasets to understand the trend of disease spreading and death. *Sensors* 2020; **20**(11): 3089.

[15]Albahli S, Algsham A, Aeraj S, Alsaeed M, Alrashed M, Rauf HT, et al. COVID-19 public sentiment insights: A text mining approach to the Gulf countries. *Comput Mater Contin* 2021; **67**(2): 1613-1627.

[16]Odhiambo J, Weke P, Ngare P. A deep learning integrated Cairns-Blake-Dowd (CBD) sytematicmortalityrisk model. *J Risk FinancManag* 2021; **14**(6): 259.

[17]Ahmad OB, Boschi-Pinto C, Lopez AD, Murray CJL, Lozano R, Inoue M. *Age standardization of rates: A new WHO standard. Geneva World Heal Organ. (Global Programme on Evidence for Health Policy Discussion Paper No. 31), 2001.* [Online]. Available from: http://www.who.int/healthinfo/paper31. pdf. [Assessed on 31 December 2017].

[18]Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE 2013. p. 6645-6649; May 2013, Vancouver, Canada.

[19]Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; **9**(8): 1735-1780.

[20]Azzouni A, Pujolle G. A long short-term memory recurrent neural network framework for network traffic matrix prediction. *arXiv:1705.05690 [cs.NI]* 2017; Prepublished.

[21]Zhao Z, Chen W, Wu X, Chen PCY, Liu J. LSTM network: A deep learning approach for short-term traffic forecast. *IET Intell Transp Syst* 2017; **11**(2): 68-75.

[22]Parsai MA. *Mazandaran COVID-19 dataset. Centers for Disease Control and Prevention*. 2020. [Online]. Available from: https://www.mazums.ac.ir. [Accessed on 10 December 2021].

[23]Ahmad S. Potential of age distribution profiles for the prediction of COVID-19 infection origin in a patient group. *Informatics Med Unlocked* 2020; **20**: 100364.

[24]Bonanad C, Garcia-Blas S, Tarazona-Santabalbina F, Sanchis J, Bertomeu-González V, Facila L, et al. The effect of age on mortality in patients with COVID-19: A meta-analysis with 611 583 subjects. *J Am Med Dir Assoc* 2020; **21**(7): 915-918.

[25]Kırbaş İ, Sözen A, Tuncer AD, Kazancıo lu F. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. *Chaos, Solitons & Fractals* 2020; **138**: 110015.

[26]Sasson I. Age and COVID-19 mortality: A comparison of Gompertz doubling time across countries and causes of death. *Demogr Res* 2021; **44**: 379-396.

[27]Ghisolfi S, Almås I, Sandefur JC, von Carnap T, Heitner J, Bold T. Predicted COVID-19 fatality rates based on age, sex, comorbidities and health system capacity. *BMJ Glob Heal* 2020; **5**(9): e003094.