Available online at www.sciencedirect.com

ScienceDirect

Biomedical Journal

journal homepage: www.elsevier.com/locate/bj

Original Article

Artificial neural network and logistic regression modelling to characterize COVID-19 infected patients in local areas of Iran

Farzaneh Mohammadi [a,b,*], *Hamidreza Pourzamani* [a,b], *Hossein Karimi* [a], *Maryam Mohammadi* [c], *Mohammad Mohammadi* [d], *Nahid Ardalan* [e], *Roya Khoshravesh* [f], *Hassan Pooresmaeil* [g], *Samaneh Shahabi* [h], *Mostafa Sabahi* [i], *Fatemeh Sadat miryonesi* [j], *Marzieh Najafi* [k], *Zeynab Yavari* [l], *Farideh Mohammadi* [m], *Hakimeh Teiri* [a,b], *Mahsa Jannati* [n]

[a] *Department of Environmental Health Engineering, School of Health, Isfahan University of Medical Sciences, Isfahan, Iran*
[b] *Environment Research Center, Research Institute for Primordial Prevention of Non-communicable Disease, Isfahan University of Medical Sciences, Isfahan, Iran*
[c] *Department of Management and Health Information Technology, School of Management and Medical Information Sciences, Isfahan University of Medical Sciences, Isfahan, Iran*
[d] *Department of Electrical Engineering, Shahreza University, Isfahan, Iran*
[e] *Kurdistan University of Medical Sciences, Sanandaj, Kurdistan, Iran*
[f] *Kermanshah University of Medical Sciences, Kermanshah, Iran*
[g] *Emergency Medical Services, Tehran, Iran*
[h] *Hamedan University of Medical Sciences, Hamedan, Iran*
[i] *Shahid Beheshti Hospital, Kashan, Isfahan, Iran*
[j] *School of Nursing & Midwifery, Isfahan University of Medical Sciences, Isfahan, Iran*
[k] *Isfahan Endocrine and Metabolism Research Center, Isfahan University of Medical Sciences, Isfahan, Iran*
[l] *Genetic and Environmental Adventures Research Center, School of Abarkouh Paramedicine, Shahid Sadoughi University of Medical Sciences, Yazd, Iran*
[m] *Department of Textile Engineering, Isfahan University of Technology, Isfahan, Iran*
[n] *Graduate Student, Dept. of Civil Engineering, Lakehead University, Thunder Bay, ON, Canada*

ARTICLE INFO

Article history:
Received 31 May 2020
Accepted 17 February 2021
Available online 25 February 2021

ABSTRACT

Background: COVID-19 is an infectious disease that started spreading globally at the end of 2019. Due to differences in patient characteristics and symptoms in different regions, in this research, a comparative study was performed on COVID-19 patients in 6 provinces of Iran. Also, multilayer perceptron (MLP) neural network and Logistic Regression (LR) models were applied for the diagnosis of COVID-19.

* *Corresponding author*. Department of Environmental Health Engineering, School of Health, Isfahan University of Medical Sciences, Hezar Jarib St., Isfahan University of Medical Sciences, Isfahan 8174673461, Iran.
E-mail address: fm_1363@hlth.mui.ac.ir (F. Mohammadi).
Peer review under responsibility of Chang Gung University.
https://doi.org/10.1016/j.bj.2021.02.006

Keywords:
COVID-19
Symptom
Epidemiology
Model
ANN
Logistic regression

Methods: A total of 1043 patients with suspected COVID-19 infection in Iran participated in this study. 29 characteristics, symptoms and underlying disease were obtained from hospitalized patients. Afterwards, we compared the obtained data between confirmed cases. Furthermore, the data was applied for building the ANN and LR models to diagnosis the infected patients by COVID-19.
Results: In 750 confirmed patients, Common symptoms were: fever (%) >37.5 °C, cough, shortness of breath, fatigue, chills and headache. The most common underlying diseases were: hypertension, diabetes, chronic obstructive pulmonary disease and coronary heart disease. Finally, the accuracy of the ANN model to the diagnosis of COVID-19 infection was higher than the LR model.
Conclusion: The prevalent symptoms and underlying diseases of COVID-19 patients were similar in different provinces, but the incidence of symptoms was significantly different from each other. Also, the study demonstrated that ANN and LR models have a high ability in the diagnosis of COVID-19 infection.

At a glance of commentary

Scientific background on the subject

Since December 2019, the coronavirus has been known as an urgent threat to global health. To help the healthcare systems, efficient diagnosis using several symptoms or features of suspected patients is essential. Until now, different Models from rule based scoring to advanced machine learning models have been proposed and published.

What this study adds to the field

Here we used artificial neural network and logistic regression to characterize COVID-19 infected patients. What distinguishes this study are the large numbers of COVID-19 suspected patients (1043) that participated in this study and also too many variables (29, demographic characteristics, symptoms and underlying disease) are included in the model.

In February 2020, the first case of coronavirus was reported in Iran. According to the latest report from the World Health Organization (WHO), the number of cases of coronavirus or COVID-19 infection in the world has reached more than 63,000,000 people and has led to the death of more than 1,466,000 people. Among these, more than 948,749 confirmed infected patients and 47,874 deaths are related to Iran (until November 30, 2020). COVID-19 with SARS and MERS is the third emerging pathogenic coronavirus for humans over the past two decades [1].

The problem that makes the Covid-19 pandemic so complicated is that it's hard to know how the virus will affect any individuals. Most people infected with the Covid-19 will present with few or mild symptoms, others may find themselves relying on a ventilator to breathe, or others die quickly. This makes it difficult to diagnose the disease based on clinical symptoms [2,3]. In the current situation, early diagnosis of coronavirus infection and timely treatment reduces its complications and spread [4]. Until now, artificial intelligence and logistical regression have been used to diagnose various diseases in many studies [5–7].

Therefore in this study, we had two main goals; first, we perform a statistical analysis and comparison on the characteristics, symptoms and underlying disease of COVID-19 patients in 6 provinces in Iran and investigate if there is a significant difference between them; second, the MLP neural network and logistic regression were used to predict binary responses in COVID-19 infection diagnosis. Afterwards, the ability of the two models was compared with some performance parameters. Finally, external validation was performed to evaluate the generalizability of the newly developed diagnostic models.

Methods

Study design and data collection

This study was supported by Isfahan University of Medical Sciences (Research Project, # 198327 and Ethic code IR.MUI.-MED.REC.1399.001.), additionally the consent form approved by the Ministry of Health of the Islamic Republic of Iran was received from all participants (both original and validation patients).

The medical records and clinical data were obtained from 1043 suspected patients with COVID-19 infection. The confirmation of COVID-19 infection was performed by Chest CT and RT-PCR testing in laboratories approved by the Iran Ministry of Health and Medical Education. Necessary data and information were extracted from questionnaires filled out by the nurses at the time of triage on Covid-19 wards from suspected patients. The hospitals under study are located in 7 provinces in Iran, as shown in [Fig. 1]. The data are divided into 6 groups. The provinces under study are Isfahan, Tehran, Kurdistan, Kermanshah, Hamedan and Chahar Mahal. Data from a hospital in Yazd province were used for external validation of the diagnostic models, but, not used in the model developing stage.

Fig. 1 Distribution of the data obtained from the 6 provinces of Iran.

The six groups of patients Compared with 29 variables which including demographic, epidemiological and clinical symptoms and characteristics of participants, those are: Age, sex, smoking (The person him/herself or his/her roommate), fever, nasal congestion, headache, cough, sore throat, sputum, runny nose, frequent sneezing, fatigue, shortness of breath, nausea or vomiting, diarrhoea, myalgia or arthralgia, chills, throat congestion, tonsil swelling, reduced sense of smell, reduced sense of taste, chronic obstructive pulmonary disease, diabetes, hypertension, coronary heart disease, cerebrovascular disease, immunodeficiency, cancer, chronic renal disease.

Statistical analysis

Continuous variables are expressed as mean ± SD and median and interquartile ranges (25th, 50th and 75th percentile) Analysis of variance (ANOVA) was used for comparing means of continuous variables in more than two independent groups, categorical variables are represented by a percentage and were compared by the χ^2 test in more than two independent groups. The Kruskal–Wallis test evaluates the differences between three or more groups in ordinal variables. In this study, the ANOVA analysis, χ^2 test and Kruskal–Wallis test were used respectively to compare the mean age, symptoms and underlying disease and Fever of confirmed patients between the studied provinces. The analyses were performed by non-missing data. The SPSS 26 statistical software was used for analysis, and p-value < 0.05 was considered statistically significant.

Modelling for diagnosis of COVID-19 infection

Logistic regression is a statistical regression model for binary dependent variables such as infection or non-infection, disease or health, death or life [8,9]. Logistic regression was implemented in SPSS 26 software. All 29 variables were entered into the LR model as independent variables. The response or dependent variable in this study is infected and not infected with COVID-19. A total of 870 COVID-19 suspected patients (638 confirmed, 232 unconfirmed) were selected to train the LR model with the Enter method and the remaining 153 patients (113 confirmed, 41 unconfirmed) were used for testing. It is necessary to note that in this study, because the data are imbalanced, the Stratified Random Sampling (SRS) method was used for training and testing sampling. Stratification will ensure that the percentages of each class in entire data will be the same (or very close to) within each individual subgroups (more details explained in suplamentary materials) [10].

MATLAB 2014 software was used to build the MLPNN model. The neural network was developed using the Neural

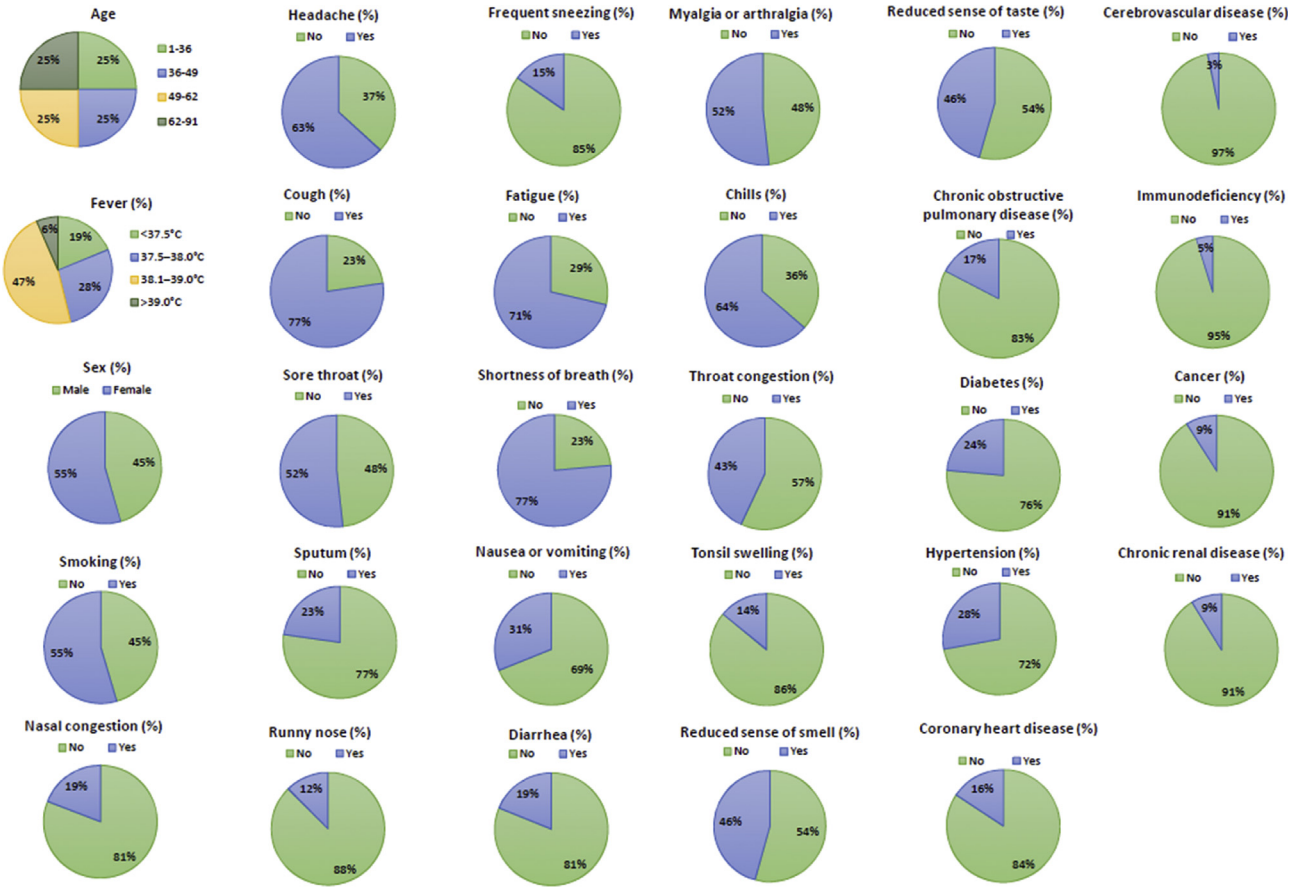

Fig. 2 Characteristics and symptoms of total confirmed Covid-19 patients in the study (n = 750).

Net Pattern Recognition toolbox (nprtool). In pattern recognition problems, the ANN used to classify inputs into a set of target categories. Here, a neural network was developed with the entry of all independent studied variables (29 variables). The neural network created includes the input layer, one hidden layer, and the output layer. A two-layer feed-forward network, with Hyperbolic tangent sigmoid and softmax activation functions in hidden and output layers, could classify vectors arbitrarily well, given enough neurons in its hidden layer. In this study, equations (1–3) were applied for determining the number of neurons in the hidden layer.

$$n_h < \frac{i + \sqrt{n}}{L} \tag{1}$$

$$\frac{2(i+o)}{3} < n_h < i(i+o) - 1 \tag{2}$$

$$0.5i - 2 < n_h < 2i + 2 \tag{3}$$

where i, o, n_h, L, n are the number of inputs neurons, number of outputs neurons, number of hidden layer neurons, number of hidden layer and number of datasets [11–13].

In next step, 717 datasets (70%) were applied for ANN training (526 confirmed, 191 unconfirmed), and the remaining one-half was used for validation (153 datasets,15%, 113 confirmed, 41 unconfirmed) and testing (153 datasets, 15%, 113 confirmed, 41 unconfirmed). The network will be trained with scaled conjugate gradient (SCG) Backpropagation algorithm. To evaluate ANN performance cross-entropy and confusion matrix was used. The predictions of both the ANN and LR models in the testing group of 153 patients were reported. Also, for external validation, information of 20 patients suspected of COVID-19 infection was received from a hospital in Yazd province and the performance of two developed diagnostic models were evaluated.

The ability and accuracy of the ANN and LR models, which are classifier models, were compared in predicting COVID-19 infected patient using the area under the receiver operating characteristic (ROC) curve. Other performance parameters were estimated using equations (4–6).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{5}$$

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{6}$$

Here, TP, FN, FP, TN, P and N are true positive, false negative, false positive, true negative, positive and negative, respectively [14].

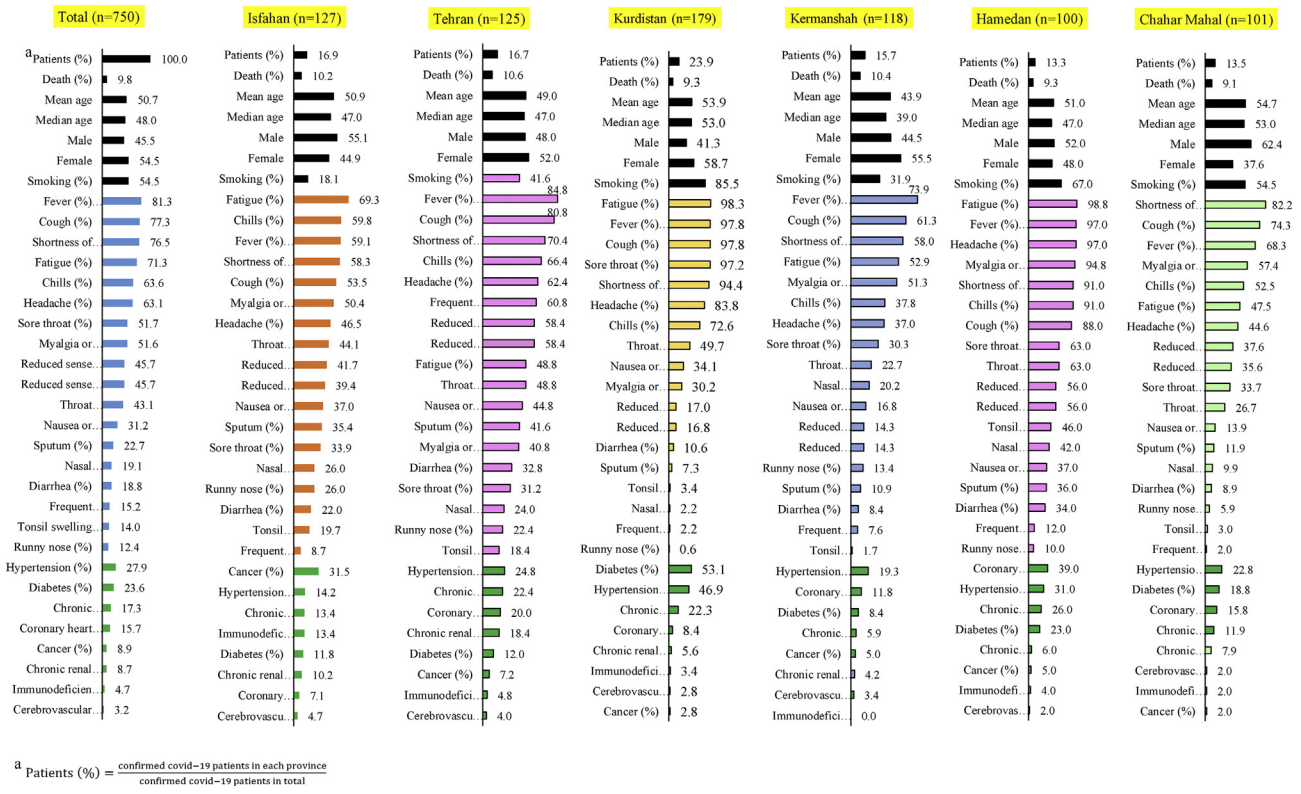

Fig. 3 Comparison of characteristics and symptoms of confirmed Covid-19 patients.

Results

Characteristics of total confirmed infected patients with COVID-19

Totally 750 of 1023 hospitalized patients was confirmed to have COVID-19 infection, those patients were selected from 12 hospitals from 6 provinces in Iran. The total data are summarized in [Table 1]. 273 (26.7%) of hospitalized patients, despite having symptoms, but they were not infected by COVID-19 and infected by other Acute Respiratory Syndromes. 57 (5.6%) confirmed patients were doctors, nurses, and other medical staff. About 558 (54.5%) of confirmed patients exposed to smoking.

Characteristics and symptoms of total confirmed Covid-19 patients in this study were plotted in [Fig. 2]. The mean and median age of confirmed patients was 50.7 ± 17.7 and 48.0 years (between 1 and 91 years, 25th, 50th and 75th percentile were 37.0, 48.0 and 63.0); only 4 (0.39%) were children below 15 years; 174 (17.0%) were 65 years old and over, and 3 pregnant women (38, 26 and 34 years old) which all were discharged safely from the hospital. 557 (54.5%) and 466 (45.5%) patients were female and male. During the study period, 74 (9.8%) of 750 patients died.

The observed symptoms of total COVID-19 patients based on [Fig. 2] were fever>37.5 °C (81.3%), cough (77.3%), shortness of breath (76.5%), fatigue (71.3%), Chills (63.6%), headache (63.1%), Sore throat (51.7%), Myalgia or arthralgia (51.6%), Reduced sense of smell (54.7%), Reduced sense of taste (45.7%), Throat congestion (43.1%), Nausea or vomiting (31.2%), Sputum (22.7%), Nasal congestion (19.1%), Diarrhea (18.8%), Frequent sneezing (15.2%) Tonsil swelling (14.0%), and Runny nose (12.4%).

The underlying disease of total COVID-19 patients according to [Fig. 2] were Hypertension (27.9%), Diabetes (23.6%), Chronic obstructive pulmonary, (17.3%) Coronary heart disease (15.7%), Cancer (8.9%), Chronic renal disease (8.7%), Immunodeficiency (4.7%), Cerebrovascular disease (3.2%). Among hospitalized patients with COVID-19, 84 (11.2%) admitted to the ICU. Also, the underlying and chronic disease was more common among patients admitted to the ICU.

Comparison of COVID-19 patients in 6 provinces of Iran

In this study, a total of 750 confirmed patients were examined in 6 provinces in Iran. The number of confirmed patients from different provinces are Isfahan 127 (16.9%), Tehran 125 (16.7%), Kurdistan 179 (23.9%), Kermanshah 118 (15.7%), Hamedan 100 (13.3%) and Chahar Mahal 101 (13.5%).

The characteristics, symptoms and underlying disease for every province are summarized in [Table 1] and [Fig. 3]. The mortality rate varies between 9.1% in Chahar Mahal to 10.6% in Tehran. There is a statistically significant difference between groups (p-value<0.05). By comparing one by one, the provinces were divided into two subgroups, Tehran, Kermanshah and Isfahan in the first group and others in another group.

Statistical analysis of characteristics, symptoms and underlying disease between 6 provinces showed the statistically

Table 1 Characteristics and symptoms of the Studied Patients.

Patients (Capita)		Total with external validation data	Total without external validation data	Isfahan	Tehran	Kurdistan	Kermanshah	Hamedan	Chahar Mahal	External validation (Yazd)
Total Patients		1043	1023	171	173	248	156	135	140	20
Confirmed Cases		762	750	127	125	179	118	100	101	12
Unconfirmed Cases		281	273	44	48	69	38	35	39	8
Variable	Total without external validation data	Confirmed Infected Patients without external validation data								Yazd (Confirmed Infected)
		Total	Isfahan	Tehran	Kurdistan	Kermanshah	Hamedan	Chahar Mahal	*p*-value[d]	
Age										
Mean	48.94 ± 18.27	50.7 ± 17.7	50.9 ± 17.9	49.0 ± 15.3	53.9 ± 16.3	43.9 ± 17.2	51.0 ± 18.8	54.7 ± 19.7	0.000[a]	60.2 ± 16.46
Median	47.0	48.0	47.0	47.0	53.0	39.0	47.0	53.0		60.0
Range	90.0	90.0	72.0	66.0	72.0	89.0	72.0	75.0		52.0
Percentile 25	36.0	37.0	38.0	38.0	40.0	32.0	38.8	36.5		47.8
Percentile 50	49.0	48.0	47.0	47.0	53.0	39.0	47.0	53.0		60.0
Percentile 75	62.0	63.0	63.0	61.0	67.0	54.0	64.8	71.0		73.5
Sex (%)										
Male	47.7	45.5	55.1	48.0	41.3	44.5	52.0	62.4	0.013[b]	58.3
Female	52.3	54.5	44.9	52.0	58.7	55.5	48.0	37.6		41.7
Fate (%)										
Death	–	9.8	10.2	10.6	9.3	10.4	9.3	9.1	0.042[b]	0
Survival	–	90.2	89.8	89.4	90.7	89.6	90.7	90.9		0
smoking (The person him/herself or his/her roommate) (%)										
No	56.8	45.5	81.9	58.4	14.5	68.1	33.0	45.5	0.000[b]	83.3
Yes	43.2	54.5	18.1	41.6	85.5	31.9	67.0	54.5		16.7
Fever (%)										
<37.5 °C	28.3	18.7	40.9	15.2	2.2	26.1	3.0	31.7	0.000[c]	25.0
37.5–38.0 °C	27.1	27.6	21.3	36.8	30.7	31.1	34.0	7.9		25.0
38.1–39.0 °C	38.6	47.2	34.6	39.2	63.7	37.8	63.0	38.6		50.0
>39.0 °C	6.0	6.5	3.1	8.8	3.4	5.0	0.0	21.8		0.0
Nasal congestion (%)										
No	81.8	80.9	74.0	76.0	97.8	79.8	58.0	90.1	0.000[b]	75.0
Yes	18.2	19.1	26.0	24.0	2.2	20.2	42.0	9.9		25.0
Headache (%)										
No	50.2	36.9	53.5	37.6	16.2	63.0	3.0	55.4	0.000[b]	66.7
Yes	49.8	63.1	46.5	62.4	83.8	37.0	97.0	44.6		33.3
Cough (%)										
No	36.5	22.7	46.5	19.2	2.2	38.7	12.0	25.7	0.000[b]	33.3
Yes	63.5	77.3	53.5	80.8	97.8	61.3	88.0	74.3		66.7
Sore throat (%)										
No	53.7	48.3	66.1	68.8	2.8	69.7	37.0	66.3	0.000[b]	41.7
Yes	46.3	51.7	33.9	31.2	97.2	30.3	63.0	33.7		58.3
Sputum (%)										
No	75.9	77.3	64.6	58.4	92.7	89.1	64.0	88.1	0.000[b]	50.0
Yes	24.1	22.7	35.4	41.6	7.3	10.9	36.0	11.9		50.0

(*continued on next page*)

Table 1 – (*continued*)

Patients (Capita)		Total with external validation data	Total without external validation data	Isfahan	Tehran	Kurdistan	Kermanshah	Hamedan	Chahar Mahal	External validation (Yazd)
Runny nose (%)										
No	75.2	87.6	74.0	77.6	99.4	86.6	90.0	94.1	0.000[b]	91.7
Yes	24.8	12.4	26.0	22.4	0.6	13.4	10.0	5.9		8.3
Frequent sneezing (%)										
No	75.6	84.8	91.3	39.2	97.8	92.4	88.0	98.0	0.000[b]	91.7
Yes	24.4	15.2	8.7	60.8	2.2	7.6	12.0	2.0		8.3
Fatigue (%)										
No	46.7	28.7	30.7	51.2	1.7	47.1	1.2	52.5	0.000[b]	8.3
Yes	53.3	71.3	69.3	48.8	98.3	52.9	98.8	47.5		91.7
Shortness of breath (%)										
No	43.4	23.5	41.7	29.6	5.6	42.0	9.0	17.8	0.000[b]	0.0
Yes	56.6	76.5	58.3	70.4	94.4	58.0	91.0	82.2		100.0
Nausea or vomiting (%)										
No	72.9	68.8	63.0	55.2	65.9	83.2	63.0	86.1	0.000[b]	58.3
Yes	27.1	31.2	37.0	44.8	34.1	16.8	37.0	13.9		41.7
Diarrhoea (%)										
No	83.4	81.2	78.0	67.2	89.4	91.6	66.0	91.1	0.000[b]	75.0
Yes	16.6	18.8	22.0	32.8	10.6	8.4	34.0	8.9		25.0
Myalgia or arthralgia (%)										
No	61.4	48.4	49.6	59.2	69.8	48.7	5.2	42.6	0.000[b]	66.7
Yes	38.6	51.6	50.4	40.8	30.2	51.3	94.8	57.4		33.3
Chills (%)										
No	50.6	36.4	40.2	33.6	27.4	62.2	9.0	47.5	0.000[b]	41.7
Yes	49.4	63.6	59.8	66.4	72.6	37.8	91.0	52.5		58.3
Throat congestion (%)										
No	60.4	56.9	55.9	51.2	50.3	77.3	37.0	73.3	0.000[b]	50.0
Yes	39.6	43.1	44.1	48.8	49.7	22.7	63.0	26.7		50.0
Tonsil swelling (%)										
No	88.4	86.0	80.3	81.6	96.6	98.3	54.0	97.0	0.000[b]	91.7
Yes	11.6	14.0	19.7	18.4	3.4	1.7	46.0	3.0		8.3
Reduced sense of smell (%)										
No	63.2	54.3	58.3	41.6	82.7	85.7	44.0	62.4	0.000[b]	41.3
Yes	36.8	45.7	41.7	58.4	17.0	14.3	56.0	37.6		58.7
Reduced sense of taste (%)										
No	63.2	54.3	60.6	41.6	83.2	85.7	44.0	64.4	0.000[b]	41.3
Yes	36.8	45.7	39.4	58.4	16.8	14.3	56.0	35.6		58.7
Chronic obstructive pulmonary disease (%)										
No	87.1	82.7	86.6	77.6	77.7	94.1	74.0	88.1	0.000[b]	91.7
Yes	12.9	17.3	13.4	22.4	22.3	5.9	26.0	11.9		8.3
Diabetes (%)										
No	82.2	76.4	88.2	88.0	46.9	91.6	77.0	81.2	0.000[b]	83.3
Yes	17.8	23.6	11.8	12.0	53.1	8.4	23.0	18.8		16.7

Hypertension (%)										
No	78.1	72.1	85.5	75.2	53.6	80.7	69.0	77.2	0.000[b]	83.3
Yes	21.9	27.9	14.2	24.8	46.9	19.3	31.0	22.8		16.7
Coronary heart disease (%)										
No	86.1	84.3	92.9	80.0	91.6	88.2	61.0	84.2	0.000[b]	83.3
Yes	13.9	15.7	7.1	20.0	8.4	11.8	39.0	15.8		16.7
Cerebrovascular disease (%)										
No	96.7	96.8	95.3	96.0	97.2	96.6	98.0	98.0	0.811[b]	91.7
Yes	3.3	3.2	4.7	4.0	2.8	3.4	2.0	2.0		8.3
Immunodeficiency (%)										
No	96.2	95.3	86.6	95.2	96.6	100.0	96.0	98.0	0.000[b]	100.0
Yes	3.8	4.7	13.4	4.8	3.4	0.0	4.0	2.0		0.0
Cancer (%)										
No	92.8	91.1	68.5	92.8	97.2	95.0	95.0	98.0	0.000[b]	100.0
Yes	7.2	8.9	31.5	7.2	2.8	5.0	5.0	2.0		0.0
Chronic renal disease (%)										
No	93.3	91.3	89.8	81.6	94.4	95.8	94.0	92.1	0.001[b]	91.7
Yes	6.7	8.7	10.2	18.4	5.6	4.2	6.0	7.9		8.3

[a] ANOVA test.
[b] Chi-Square Tests.
[c] Kruskal Wallis Test.
[d] The *p*-values determine if there is a significant difference between the variables in different provinces in confirmed Covid-19 patients, *p*-value less than 0.05 is statistically significant.

significant difference between the groups (*p*-value< 0.01) except cerebrovascular disease which was generally the least common among other underlying diseases.

In each province, the 5 top common symptoms were different, as follow [Fig. 3]; **Isfahan**: fatigue (69.3%), chills (59.8%), fever (59.1%), shortness of breath (58.3%), and cough (53.5%). **Tehran**: fever (84.8%), cough (80.8%), shortness of breath (70.4%), chills (66.4%) and headache (62.4%). **Kurdistan**: fatigue (98.3%), fever (97.8%), cough (97.8%), sore throat (97.2%) and shortness of breath (94.4%). **Kermanshah**: fever (73.9%), cough (61.3%), shortness of breath (58.0%), fatigue (52.9%) and myalgia or arthralgia (51.3%). **Hamedan**: fatigue (98.8%), fever (97.0%), Headache (97.0%), myalgia or arthralgia (94.8%) and shortness of breath (91.0%). **Chahar Mahal**: shortness of breath (82.2%), cough (74.3%), fever (68.3%), myalgia or arthralgia (57.4%) and chills (52.5%).

The most common underlying disease observed in each province was as follow; **Isfahan**, cancer (31.5%), **Tehran**, hypertension (24.8%), **Kurdistan**, diabetes (53.1%), **Kermanshah**, hypertension (19.3%), **Hamedan**, Coronary heart disease (39.0%) and **Chahar Mahal**, hypertension (22.8%). In Isfahan province, cancer has been observed in 31.5% of COVID-19 patients, that's why one of the investigated hospitals was especially for cancer patients.

From the results of statistical analysis performed in [Table 1] and [Figs. 2 and 3], it is concluded that the prevalent symptoms and underlying diseases of COVID-19 patients were similar in different provinces, but the incidence of symptoms was significantly different from each other.

ANN and LR models for COVID-19 disease diagnosis

29 independent variables of 1023 total suspected patients were used to build the ANN and LR models. The output classes were not-infected and infected by Covid-19. The Omnibus Tests indicates that the accuracy of the LR model improves when the variables added to the model (*p*-values< 0.001). Cox & Snell R Square and Nagelkerke R Square are equal to 0.683 and 0.965 respectively and indicating a strong relationship between the predictors and the prediction. In Hosmer and Leme test unlike most, *p*-values should be more than 0.05 to indicate a good fit to the data and in this study the p-value = 1.00 so the LR model is reliable. The accuracy of logistic regression classification for training datasets was 98.9% (n = 870). [Table 2] provides the Wald statistic which is significant if the *p*-value<0.05. In this study, the presence of 16 variables in the equation was significant. Thus, age, sex, smoking, nasal congestion, sputum, tonsil swelling, diarrhoea, chronic obstructive pulmonary disease, diabetes, hypertension, coronary heart disease, cerebrovascular disease and chronic renal disease were removed from the equation.

In the ANN model, the number of neurons in hidden layer based on equations (1)–(3) was assessed in 20–60 interval by trial and error. 20 neurons in the hidden layer showed the best performance. The ANN structure is shown in [Fig. 4-A]. The best performance of optimized ANN based on cross-entropy in training, validation and test steps is visible in [Fig. 4-B]. The minimum cross-entropy occurred in epoch 11 and equal to 0.1077.

[Fig. 5] and [Table 3] demonstrate the high ability of both models in the diagnosis of COVID-19. For ANN model, the area

Table 2 Variables in the Equation based on the LR model.

Variables	Wald	df	*p*-value
Fever	13.140	3	0.004
Shortness of Breath	28.759	1	0.000
Headache	4.290	1	0.038
Cough	12.342	1	0.000
Fatigue	24.451	1	0.000
Chills	17.455	1	0.000
Sore Throat	4.650	1	0.031
Myalgia or Arthralgia	24.275	1	0.000
Runny Nose	22.143	1	0.000
Frequent Sneezing	25.167	1	0.000
Reduced Sense of Smell	5.719	1	0.017
Reduced Sense of Taste	8.352	1	0.004
Nausea or vomiting	4.965	1	0.026
Throat congestion	5.022	1	0.025
Immunodeficiency	8.185	1	0.004
Cancer	6.135	1	0.013
Constant	24.329	1	0.000

under the ROC curve (AUC) was 0.999 (95% confidence interval = 0.998–1.0, p-value<0.05) which was higher than of LR model with AUC = 0.992 (95% confidence interval = 0.987–0.998, P-value < 0.05). The ANN model had a sensitivity of 100.0%, a specificity of 97.6% and an accuracy of 99.4%. The LR model had a sensitivity of 99.1%, a specificity of 97.6% and an accuracy of 98.7%. The ANN and LR models were evaluated on the testing group of 153 patients. The confusion matrix for these data was shown in [Fig. 6] Based on the mentioned parameters, the ANN model was better performance than the LR model.

Prediction models tended to perform better on data that models were constructed than on new data. This highlights the importance of external validation. In this research, due to the limitations of internal validation to determine the generalizability of diagnostic prediction models, the external validation was performed [15,16]. For this purpose, information of 20 patients suspected to COVID-19 was collected from a hospital in Yazd province. The data of these patients were considered as new for both diagnostic models. The simulation results were very interesting. As [Fig. 7] shows, the ANN model can correctly predict infected and not-infected patients 100%. The LR model also performed very well and only it misdiagnosed one person, in a way that a not-infected patient was diagnosed as infected. Also, For external validation data the AUC, sensitivity, specificity and accuracy of the diagnostic models could be seen in [Table 3].

Discussion

Severe Acute Respiratory Syndrome (SARS-CoV-2) is a new strain of coronavirus that has not been previously identified

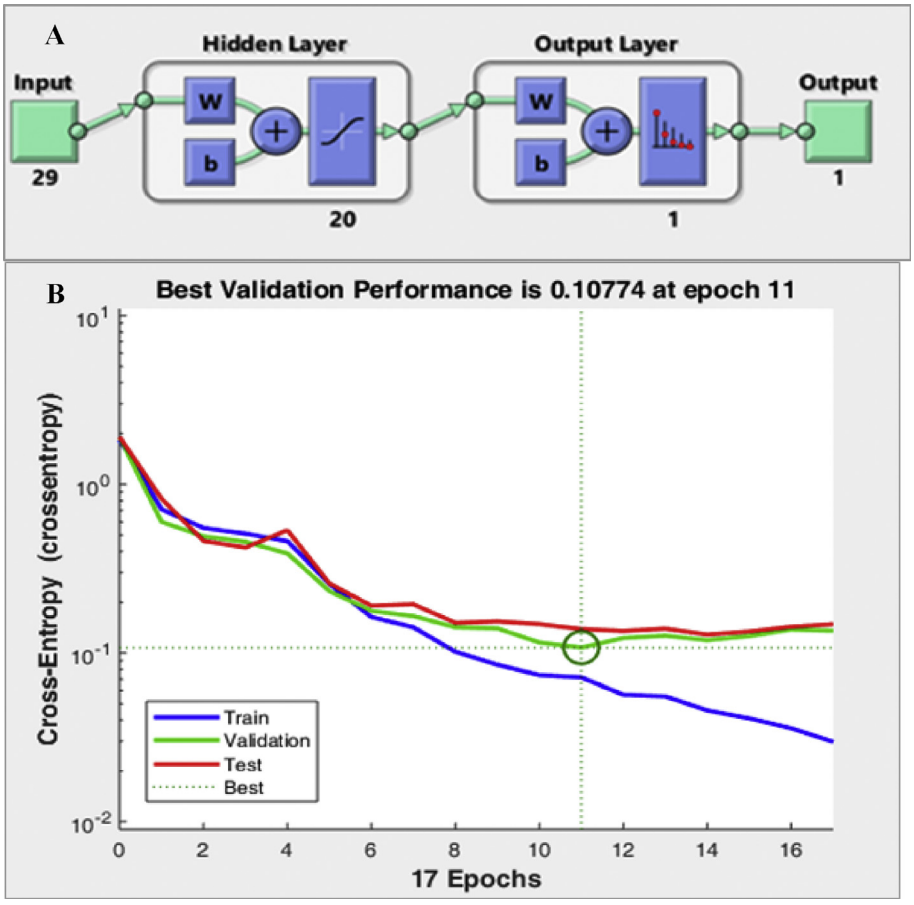

Fig. 4 A- The structure of optimized ANN, B-The performance graph of optimized ANN model to diagnose the Covid-19 infection using 29 variables determined in this study.

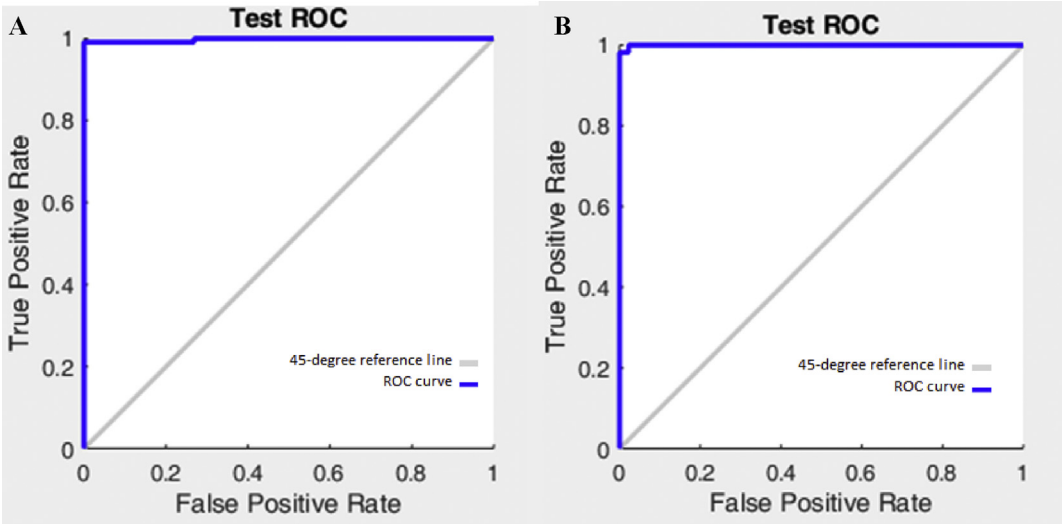

Fig. 5 The ROC curves of A- ANN model and B- LR model to diagnose the Covid-19 infection using 29 variables determined in this study.

Table 3 The performance parameters of the LR and ANN model for test data and External validation data.

Model		LR	ANN
Test data			
AUC		0.992	0.999
Asymptotic Sig		0.000	0.000
Asymptotic 95% Confidence Interval	Lower Bound	0.987	0.998
	Upper Bound	0.998	1.000
Sensitivity		0.991	1.000
Specificity		0.976	0.976
Accuracy		0.987	0.994
External validation data			
AUC		0.971	1.000
Asymptotic Sig		0.000	0.000
Asymptotic 95% Confidence Interval	Lower Bound	0.917	1.000
	Upper Bound	1.000	1.000
Sensitivity		1.000	1.000
Specificity		0.875	1.000
Accuracy		0.950	1.000

in humans. Mortality of COVID-19 appears to be higher than influenza and lower than SARS and MERS [17]. This study investigated the characteristics, symptoms and underlying diseases of COVID-19 patients in 6 provinces of Iran and compared them to know if these cases are significantly different. Although the epidemic prediction is essential for applying effective prevention and control of infectious diseases [7], it has been somewhat neglected in research for COVID-19 by now. Hence, using data obtained from hospitalized suspected COVID-19 patients, the ANN and LR models were developed for diagnostics of COVID-19-infected and not-infected patients. The age of patients was from 1 to 91 years old, and about 17.0% of patients were over 65 years of age. There was no significant difference between male and female at the 0.05 level.

Based on this study in Iran, only about 20% of those admitted to hospitals due to COVID-19 are hospitalized, and among them, approximately 8.5% are admitted to the ICU. An average of 9.8% mortality rate was calculated among hospitalized patients, therefore, the total mortality rate would be about 1.96%. In this research, severe symptoms in older, obese and overweight patients were significantly more than other patients. Mortality rates were significantly higher in elderly patients over 65 years old [18]. The mean age of died patients was 66.4 ± 16.7 years (between 22 and 90 years).

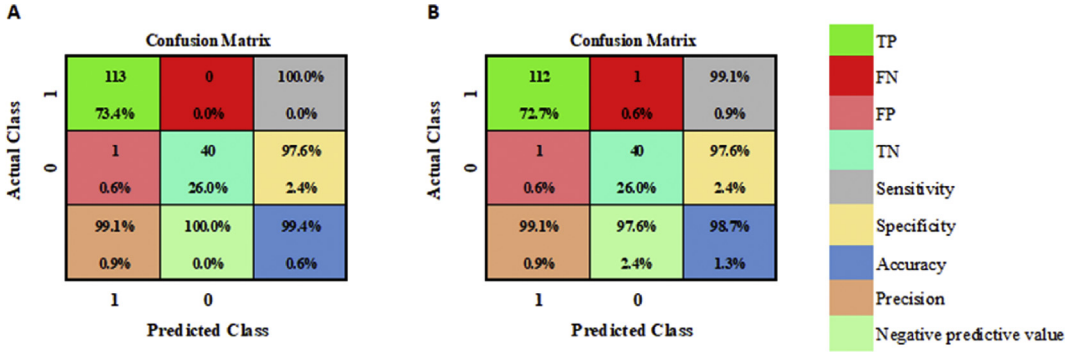

Fig. 6 The Confusion Matrix of A-ANN and B-LR model for the test dataset to diagnose the Covid-19 infection using 29 variables determined in this study.

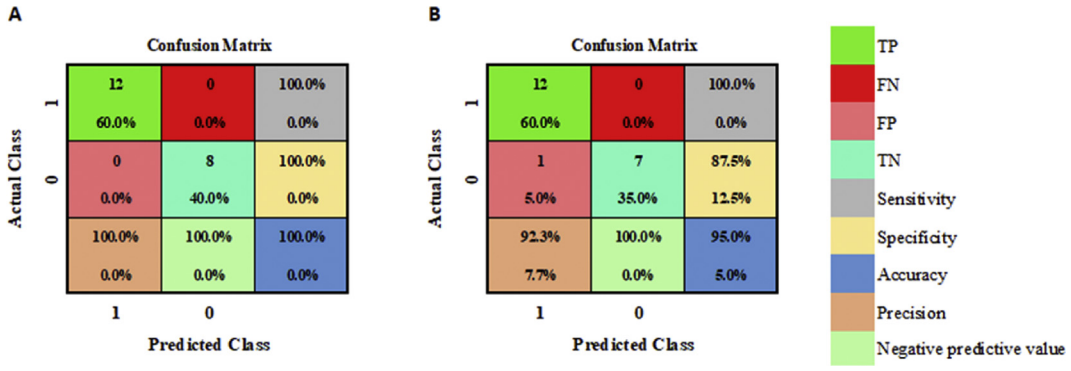

Fig. 7 The External Validation of A-ANN and B-LR models for the Yazd province patients to diagnose the Covid-19 infection using 29 variables determined in this study.

Also, patients with underlying heart disease might be more likely in the risk of severe infection and death.

The mortality rate in Tehran and Isfahan, industrialized and more populous provinces, was higher than the others. They are often heavily involved in environmental issues such as air pollution and pulmonary and heart diseases have a higher rate in these provinces [19].

The results of this study indicated that the symptoms of Covid-19 are a little different from those of SARS-CoV. The dominant symptoms in SARS are fever and cough and gastrointestinal symptoms were uncommon [20], but dominant symptoms in COVID-19 are fever, cough, shortness of breath, fatigue, chills, headache, sore throat and myalgia or arthralgia which were observed in more than 50% of patients. The gastrointestinal symptoms in COVID-19 such as nausea or vomiting and diarrhoea observed in 31.2% and 18.8% of patients, respectively.

In Isfahan, Kurdistan and Hamedan, fatigue, and in Chahar Mahal, shortness of breath, and in Tehran and Kermanshah, fever was predominant. In Isfahan, Tehran, Kurdistan and Hamedan nausea or vomiting was observed in approximately 40% and diarrhoea in Tehran and Hamedan was observed in about 35% of patients. But it is important to note that the symptoms of COVID-19 are more similar to MERS-CoV infection. Because most confirmed MERS-CoV cases have had fever, cough, shortness of breath and some others also had nausea and vomiting and diarrhoea [21]. Most common underlying disease among MERS-CoV patients are diabetes, cancer, chronic lung disease, chronic heart disease and chronic kidney disease [20] and the most common underlying disease among COVID-19 patients in this study were hypertension, diabetic, chronic obstructive pulmonary disease, coronary heart disease, cancer, chronic renal disease.

In Isfahan province, where more than a third of the people evaluated were cancer patients, fewer symptoms were observed. Among 40 cancer patients, the most common symptoms were chills (61.1%), fatigue (55.6%), fever>38°c (55.6%), nausea or vomiting (50%), shortness of breath (44.4%), throat congestion (44.4%), sputum (40.0%), cough (33.3%), myalgia or arthralgia (27.8%), headache (22.2%), sore throat (22.2%), diarrhoea (16.8%) and except cancer the another underlying disease were immunodeficiency (38.9%), chronic obstructive pulmonary disease (22.2%), diabetes (16.7%) and hypertension (11.1%).

Due to limited laboratory diagnostic testing, there were no reliable data on the prevalence of the COVID-19 virus in different population. So, methods that accelerate the diagnosis and allow for screening of the people, especially for areas with a shortage of health care worker, could be very efficient. Considering the highly contagious nature and high prevalence of COVID-19, model development for the diagnosis of COVID-19 is considered to be a crucial measure for the control of the disease. Many studies have applied the multilayer perceptron neural network and logistic regression in the diagnosis of infectious disease [7]. But no studies have compared the abilities of ANN and LR models to predict the COVID-19 infection.

In this study, the ANN and LR models were applied to predict and diagnose COVID-19 Infection. Then, the ability of models by AUC, sensitivity, specificity and accuracy were compared to classify infected (750) and not-infected patients (273). We built these models with 29 obtained variables including characteristics, symptoms and underlying disease of 1023 hospitalized patients to help patient classification and clinical decision making in the absence of standardized tests for COVID-19 Infection. Finally, external validation for the new diagnostic model was developed to verify its generalizability. The results of this study demonstrated that both the ANN and LR models were performed well, however, the ANN model achieved superior performance compared to the LR model but the difference was not significant. A meta-analysis study investigated 28 articles and revealed that ANN in 36% and LR in 14% of studies performed with higher prediction accuracy, and in other studies (50%) both models show similar performance [22].

It should be noted that, in published articles that used mathematical and machine learning models to diagnose Covid-19 patients, either the number of data was much less than this study, or if the data were extensive, the variables evaluated were much less than this study. Xiong et al.

investigated Pseudo-likelihood based logistic regression for estimating COVID-19 infection and case fatality rates by gender, race, and age in California. Their model was focused on the gender, race, and age parameters and they have not introduced the symptoms of patients to the model. Their analysis indicates that in California, males had higher infection and case fatality rates across age and race groups. Elderly infected with COVID-19 were at an elevated risk of mortality. LatinX and African Americans had higher infection rates than other race groups [15].

Machine learning-based approaches have been investigated by Khanday et al. for detecting COVID-19 using clinical text data. They used 212 clinical reports which were labelled in four classes namely COVID, SARS, ARDS and both (COVID, ARDS). Various features like TF/IDF, a bag of words were extracted from these clinical reports. The machine learning algorithms were used for classifying clinical reports into four different classes. After performing classification, it was revealed that logistic regression and multinomial Naïve Bayesian classifier gives excellent results by having 96.2% accuracy. They expressed that the efficiency of models can be improved by increasing the amount of data [23].

Shaban et al. detected COVID-19 patients based on fuzzy inference engine and Deep Neural Network. Patients' laboratory findings were introduced to the model. The total number of cases in this study was 279 (177 confirmed and 102 unconfirmed) [24]. In some other studies, Mathematical and computational models which are epidemiological models have been used to predict the number of cases of COVID-19 and infection rates [25–27].

The strengths of our study were making full use of demographical and clinical data which is very convenient and easy to obtain to build models to predict the confirmed patients. Our models help make more accurate detection of COVID-19, thus optimizing patient selection for appropriate treatment. In addition, the entry of information from more than a thousand people from different regions has greatly increased the accuracy of the model in COVID-19 detecting. However, this study has some limitations as well, such as some parts of the data received were through self-declaration of participants for determining whether the participants are infected or not with Covid-19. Also it was not possible to follow up some patients until they were discharged from the hospital.

Funding sources

This article is the result of a research project approved in the Isfahan University of Medical Sciences (IUMS), Research Project, # 198327 and Ethic code IR.MUI.MED.REC.1399.001.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgement

The authors wish to acknowledge the Vice Chancellery of Research of IUMS for the financial support and also we would like to thank all the officials, nurses and other people who helped us with this research project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bj.2021.02.006.

REFERENCES

[1] Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun 2020;109:102433.
[2] Guan WJ, Ni Z-Y, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020;382:1708–20.
[3] Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese center for disease control and prevention. J Am Med Assoc 2020;323:1239–42.
[4] Cortegiani A, Ingoglia G, Ippolito M, Giarratano A, Einav S. A systematic review on the efficacy and safety of chloroquine for the treatment of COVID-19. J Crit Care 2020;57:279–83.
[5] Das N, Topalovic M, Janssens W. Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential. Curr Opin Pulm Med 2018;24:117–23.
[6] Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, et al. Machine Learning techniques in breast cancer prognosis prediction: a primary evaluation. Cancer Med 2020;9:3234–43.
[7] Manliura Datilo P, Ismail Z, Dare J. A review of epidemic forecasting using artificial neural networks. Int J Epidemiol Res 2019;6:132–43.
[8] Brydon H, Blignaut R, Jacobs J. A weighted bootstrap approach to logistic regression modelling in identifying risk behaviours associated with sexual activity. SAHARA J 2019;16:62–9.
[9] Amin MM, Bina B, Ebrahimi A, Yavari Z, Mohammadi F, Rahimi S. The occurrence, fate, and distribution of natural and synthetic hormones in different types of wastewater treatment plants in Iran. Chin J Chem Eng 2018;26:1132–9.
[10] Ramezan C A, Warner T A, Maxwell A E. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. Rem Sens 2019;11:185.
[11] García-Alba J, Bárcena JF, Ugarteburu C, García A. Artificial neural networks as emulators of process-based models to analyse bathing water quality in estuaries. Water Res 2019;150:283–95.
[12] Ghezelbash A, Keynia F. Design and implementation of artificial neural network system for stock exchange prediction. Afr J Comp ICT 2014;7:153–60.

[13] Long J, Xueyuan K, Haihong H, Zhinian Q, Yehong W. Study on the overfitting of the artificial neural network forecasting model *. Acta Meteorol Sin 2004;19:216–25.
[14] Tong Z, Liu Y, Ma H, Zhang J, Lin B, Bao X, et al. Development, validation and comparison of artificial neural network models and logistic regression models predicting survival of unresectable pancreatic cancer. Front Bioeng Biotechnol 2020;8:196.
[15] Ing EB, Miller NR, Nguyen A, Su W, Bursztyn LLCD, Poole M, et al. Neural network and logistic regression diagnostic prediction models for giant cell arteritis: development and validation. Clin Ophthalmol 2019;13:421–30.
[16] Mohammadi F, Bina B, Amin MM, Pourzamani HR, Yavari Z, Shams MR. Evaluation of the effects of AlkylPhenolic compounds on kinetic parameters in a moving bed biofilm reactor. Can J Chem Eng 2018;96:1762–9.
[17] Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E. COVID-19, SARS and MERS: are they closely related? Clin Microbiol Infect 2020;26:729–34.
[18] Niu S, Tian S, Lou J, Kang X, Zhang L, Lian H, et al. Clinical characteristics of older patients infected with COVID-19: a descriptive study. Arch Gerontol Geriatr 2020;89:104058.
[19] Yaghoubi A, Tabrizi J-S, Mirinazhad M-M, Azami S, Naghavi-Behzad M, Ghojazadeh M. Quality of life in cardiovascular patients in Iran and factors affecting it: a systematic review. J Cardiovasc Thorac Res 2012;4:95–101.
[20] Sokouti M, Sadeghi R, Pashazadeh S, Eslami S, Sokouti M, Ghojazadeh M, et al. Comparative global epidemiological investigation of SARS-CoV-2 and SARS-CoV diseases using meta-MUMS tool through incidence, mortality, and recovery rates. Arch Med Res 2020;51:458–63.
[21] Al Sulayyim HJ, Khorshid SM, Al Moummar SH. Demographic, clinical, and outcomes of confirmed cases of Middle East Respiratory Syndrome coronavirus (MERS-CoV) in Najran, Kingdom of Saudi Arabia (KSA); A retrospective record based study. J Infect Public Health 2020;13:1342–6.
[22] Teshnizi SH, Ayatollahi SMT. A comparison of logistic regression model and artificial neural networks in predicting of student's academic failure. Acta Inf Med 2015;23:296–300.
[23] Khanday AMUD, Rabani ST, Khan QR, Rouf N, Mohi Ud Din M. Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 2020;12:731–9.
[24] Shaban WM, Rabie AH, Saleh AI, Abo-Elsoud MA. Detecting COVID-19 patients based on fuzzy inference engine and Deep Neural Network. Appl Soft Comput 2020;99:106906.
[25] Torrealba-Rodriguez O, Conde-Gutiérrez RA, Hernández-Javier AL. Modeling and prediction of COVID-19 in Mexico applying mathematical and computational models. Chaos, Solit Fractals 2020;138:109946.
[26] Salgotra R, Gandomi M, Gandomi AH. Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. Chaos, Solit Fractals 2020;138:109945.
[27] Dandekar R, Rackauckas C, Barbastathis G. A machine learning-aided global diagnostic and comparative tool to assess effect of quarantine control in COVID-19 spread. Patterns 2020;1:100145.