European Review for Medical and Pharmacological Sciences 2021; 25: 5556-5560

Evaluation of the prediction of CoVID-19 recovered and unrecovered cases using symptoms and patient's meta data based on support vector machine, neural network, CHAID and QUEST Models

D. AL-NAJJAR[1], H. AL-NAJJAR[2], N. AL-ROUSAN[2]

[1]Finance and Banking Sciences, Applied Science Private University, Amman, Jordan
[2]Department of Computer Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, Istanbul, Turkey

Abstract. – OBJECTIVE: This paper aims to develop four prediction models for recovered and unrecovered cases using descriptive data of patients and symptoms of CoVID-19 patients. The developed prediction models aim to extract the important variables in predicting recovered cases by using the binary values for recovered cases.

MATERIALS AND METHODS: The data were collected from different countries all over the world. The input of the prediction model contains 28 symptoms and four variables of the patient's information. Symptoms of COVID-19 include a high fever, low fever, sore throat, cough, and so on, where patient metadata includes Province, county, sex, and age. The dataset contains 1254 patients with 664 recovered cases. To develop prediction models, four models are used including neural network, support vector machine, CHAID, and QUEST models. To develop prediction models, the dataset is divided into train and test datasets with splitting ratios equal to 70%, and 30%, respectively.

RESULTS: The results showed that the neural network model is the most effective model in developing COVID-19 prediction with the highest performance metrics using train and test datasets. The results found that recovered cases are associated with the place of the patients mainly, province of the patient. Besides the results showed that high fever is not strongly associated with recovered cases, where cough and low fever are strongly associated with recovered cases. In addition, the country, sex, and age of the patients have higher importance than other patient's symptoms in COVID-19 development.

CONCLUSIONS: The results revealed that the prediction models of the recovered COVID-19 cases can be effectively predicted using patient characteristics and symptoms, besides the neural network model is the most effective model to create a COVID -19 prediction model. Finally, the research provides empirical evidence that recovered cases of COVID-19 are closely related to patients' provinces.

Key Words:
Epidemiology, Symptoms, Infection, COVID-19, Machine learning.

Introduction

COVID-19 is spreading worldwide. Many hospitals and health institutes have reported a range of symptoms, along with patient information to extract general information about the COVID-19 virus. Many researchers have recorded the main effects of COVID-19 on the patient's body by reporting symptoms, allergies, recovered and deceased cases, along with various information regarding patients' history[1-3]. Researchers found that age, sex, province, and the reason for infection were the dominant factors in determining recovered and deceased cases. In addition, several researchers have recorded specific symptoms (such as fever, cough, sore throat) that could indicate COVID-19 infection as discussed by the authors[1] and the Centers for Disease Control and Prevention.

So far, different researchers have studied different factors that can affect CoVID-19 patients around the world such as symptoms of patients[1], descriptive information of patients[2], and causes of weather[4] (i.e., low temperature, humidity,

Corresponding Author: Nadia AL-Rousan, MD; e-mail: nadia.rousan@yahoo.com

pressure, and short-wave radiation). The collected results concluded that COVID-19 can be affected by various reasons not only biological reasons as discussed in other studies[5,6]. Researchers have proposed different models for predicting COVID-19 deceased or recovered cases. Vetrugno et al[7] (2021) developed a decision tree model for confirmed and unconfirmed COVID-19 to predict the need for hospitalization or home monitoring. The results showed that the model is effective in distinguishing between confirmed and unconfirmed cases of COVID-19. Singh et al[8] (2020) proposed a prediction model for confirmed, deceased, and recovered cases using a support vector machine. The results showed that the support vector machine model is efficient in predicting COVID-19 cases using different scenarios.

Niazkar et al[9] (2020) proposed a neural network model to estimate the confirmed case of COVID-19 in various countries including China, Singapore, Japan, Singapore, Iran, Italy, South Africa, and the United States. The results indicated that to improve the prediction of COVID-19, the maximum incubation period should be included in the development of the prediction of COVID-19. In addition, authors[2] developed a recovered and deceased prediction model based on a neural network. The developed model showed that it is efficient to be used in predicting the COVID-19 cases.

After analyzing and reviewing various research works in different fields as shown in[10]. This study adopted four prediction models to distinguish between unrecovered and recovered COVID-19 cases in different countries worldwide. To achieve the aim of this study, different variables have been used to distinguish between recovered and unrecovered cases. The variables are mainly divided into two categories, including meta information of the patients and symptoms.

Materials and Methods

This research aims to develop four prediction models for recovered cases using COVID-19 patient symptoms and patient metadata (i.e., age, province, and sex). The prediction models are neural network (NN), support vector machine (SVM), QUEST, and CHAID models. The last two models are used to build a tree model based on different growing methods namely, Quest and CHAID methods. The parameters of each model are determined after various experimental tests to find the most accurate and best results for each model. The study considered 28 symptoms as discussed in[1], besides, four parameters including sex, province, country, and age. The symptoms are considered after analyzing the most recorded symptoms in the dataset. The symptoms are including high fever, low fever, cough, sputum, asymptomatic, pneumonitis, chills, chest, diarrhea, sore throat, fatigue, pneumonia, discomfort, nausea, runny nose, weak, headache, dry, myalgia, malaise, anorexia, muscle and joint pain, pharynx, vomiting, nasal problem, breathing difficulty, and dyspnea.

Data collected between January 4, 2020, and March 1, 2020, from the Korea Centers for Disease Control and Prevention (KCDC). The collected data set is analyzed, many data is deleted from the original data set for validation, and outlier data is removed. The dataset contains different patients from different countries with total number of recovered cases equal to 664 and the total number of data is 1254. In addition, deceased prediction models were omitted from this study as the number of deceased cases is 27, which is very small compared to the total number of cases, besides, the developed models failed to achieve an acceptable prediction ratio because prediction models are unable to establish a relationship with the input variables.

The collected data are filtered and tested using different statistical analysis, to select the most important and affected variables in the recovered cases. Prediction models are developed by dividing the data set into two subsets including train and test data sets. The percentages for the train and test data sets are 70% and 30%, respectively. The split ratio of the prediction models is selected after reviewing various literature studies that have proven the efficiency of the developed prediction model. Next, the performance of the prediction models is calculated by considering various performance measures including accuracy, accuracy, recall, false omission (FOR) rate and F1 score. The performance metrics are calculated after finding True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) of all the models. The performance metrics are calculated as follows:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (2)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (3)$$

$$\text{False omision rate (FOR)} = \frac{TN}{(TN + FN)} \quad (4)$$

$$\text{F1 score} = 2^* \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

where True Positives (TP): Number of recovered cases that predicted as true recovered. True Negatives (TN): Number of unrecovered cases that predicted as unrecovered. False Positives (FP): Number of recovered cases that predicted as unrecovered. False Negatives (FN): Number of unrecovered cases that predicted as recovered.

Results

The first step in data analysis is to clean the data set by removing the outlier data, afterward, the collected data is divided into train and test datasets to develop prediction models based on the input variables and selected model. Table I shows the developed CoVID-19 prediction analysis for recovered cases. The total number of train and test data was 878, and 375 with only 375 and 191 recovered cases, respectively. The developed prediction models using the train dataset showed that the accuracy values are 0.84, 0.83, 0.82, and 0.72 for SVM, NN, CHAID, and Quest models, where the highest precision, recall, FOR, and F1 score are achieved using SVM, Quest, NN, and SVM, respectively. The highest values for precision, recall, FOR, and F1 scores are equal to 0.87, 0.89, 0.82, and 0.42, respectively. Finally, the train results showed that the support vector machine is the best model to be used to train the recovered cases of COVID-19.

In addition, the test dataset showed that the accuracy values are 0.79, 0.81, 0.78, and 0.69 for SVM, NN, CHAID, and Quest models, where the highest precision, recall, FOR, and F1 score are achieved using NN, NN, Quest, SVM, respectively. The results showed that the neural network achieved the highest values for precision, recall, and F1 score, where the Quest model achieved the highest FOR value. The highest precision, recall, FOR, and F1 score are equal to 0.81, 0.82, 0.79, and 0.40, respectively. Therefore, the test results showed that the neural network model is the best model to be used to test the recovered cases of COVID-19.

The results indicated that the neural network model is capable to predict the COVID-19 recovered cases using patients' metadata and patients' symptoms. Moreover, to validate the importance of each variable in predicting the recovered cases, an importance analysis test is used as shown in Figure 1.

The importance variables analysis of the recovered cases showed that the developed models returned different importance variables as shown in Figure 1. The seven top importance variables for SVM (from most important to least important) are Province, Low_Fever, Cough, Sex, Muscle_Pain, Chest_Pain, Pneumonia. For the Neural network, the important variables are Province, Cough, Country, Low Fever, Sex, Age, Muscle Pain, where for the CHAID model, the important variables are Country, Province, Cough, Sex, Low Fever, Age, and Muscle Pain, finally Quest model showed that the top variables are Province, Low Fever, Country, Age, Cough, Sex, and Muscle Pain. The important variables of all models are ranged from 3.09% and 3.36%. The results

Table I. Classification of the recovered cases based on four prediction models.

	Models	Accuracy	Precision	Recall	FOR	F1_score
Training	SVM	0.84	0.87	0.82	0.80	0.42
	NN	0.83	0.84	0.85	0.82	0.42
	CHAID	0.82	0.84	0.82	0.80	0.41
	Quest	0.72	0.69	0.89	0.81	0.39
Testing	SVM	0.79	0.81	0.77	0.77	0.39
	NN	0.81	0.82	0.80	0.79	0.40
	CHAID	0.78	0.79	0.79	0.78	0.39
	Quest	0.69	0.64	0.87	0.78	0.37

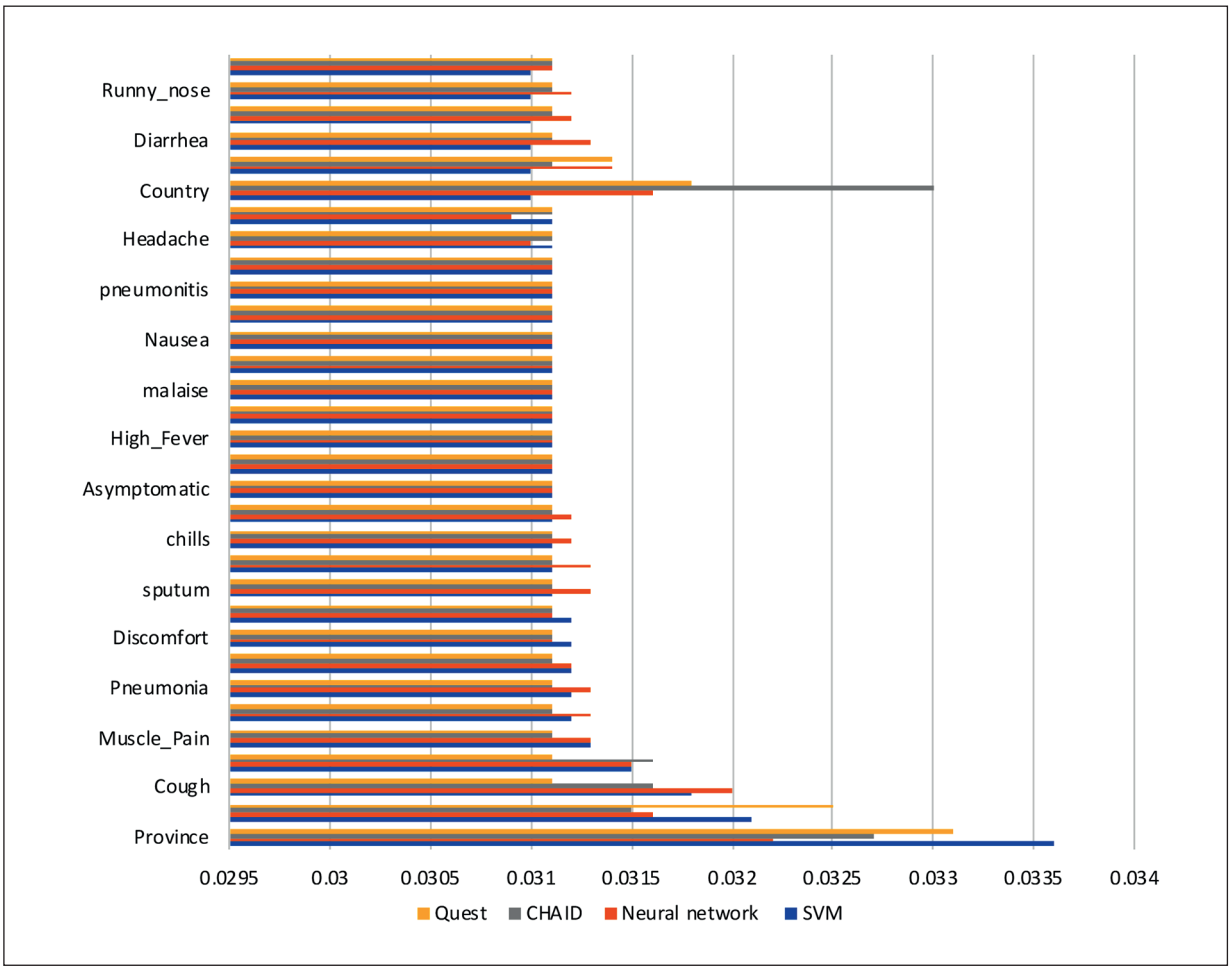

Figure 1. Importance variables for recovered cases based on four prediction models including neural network, CHAID, SVM, and Quest.

showed that the patients' Province is more important than the symptoms, country, age, and sex in all the developed prediction models.

Discussion

The results analysis revealed that symptoms are not enough to determine the recovered cases, therefore COVID-19 patients with symptoms only are not sufficient to predict the status of the patients that admitted to the hospital. Moreover, the results of the train and test data sets indicated that the neural network model is the most efficient model to be used to develop COVID-19 recovered cases. Depending on the last finding, the results indicated that the patient's metadata including Province, Country, Sex, and Age is more important than all the symptoms except low fever and cough. In addition, the low fever, and cough are the dominant variables reported in the recovered cases, where the rest of the symptoms are less important in developing a COVID-19 prediction.

Finally, recovered predictors showed that using patients' metadata had a strong influence in creating the COVID-19 prediction model. The results indicated that high fever (more than 38) was not strongly associated with recovered cases and low fever is highly correlated with the developed recovered prediction model. In addition, the results indicated that the COVID-19 viruses differ greatly from place to place. Besides, the results indicated that the virus is stronger in some provinces, and in other provinces the virus is weaker. These results are in line with the outcomes of many research[1-3], besides the analysis showed that the use of symptoms to develop a prediction is not accurate, therefore patients' metadata is important to

determine the patients' status. The results are in line with the findings of authors in[6], which found that recovery rates for CoVID-19 can vary based on the studied area.

Conclusions

This research aimed to develop different prediction models for COVID-19 recovered cases using patients' metadata and symptoms. The study used four prediction models mainly CHAID, support vector machine, neural network, and QUEST models, where the data is collected from different countries. The input variables for all the prediction models are divided into metadata of the COVID-19's patients and their symptoms. After analyzing the results obtained, it was found that the neural network model is more efficient in predicting COVID-19 recovered cases using patient symptoms with patient metadata. Moreover, the results indicated that province, cough, country, low fever, sex, and age are the most effective variables in developing COVID-19 prediction. Besides, the results find that different places may show different behaviors that can give a preliminary indication of how different COVID-19 viruses behave in different places. The results give an indirect indication that different COVID-19 mutations in different places have different characteristics. As a future study, extra deep learning analysis should be used with extra genetic variables to verify the results obtained.

Conflict of Interest

The Authors declare that they have no conflict of interests.

References

1) Al-Najjar D, Al-Najjar H, Al-Rousan N. CoVID-19 symptoms analysis of deceased and recovered cases using Chi-square test. Eur Rev Med Pharmacol Sci 2020; 24: 11428-11431.
2) Al-Rousan N, Al-Najjar H. Data Analysis of Coronavirus CoVID-19 Epidemic in South Korea Based on Recovered and Death Cases. J Med Virol 2020; 92: 1603-1608.
3) Al-Najjar H, Al-Rousan N. A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea. Eur Rev Med Pharmacol Sci 2020; 24: 3400-3403.
4) Al-Rousan N, Al-Najjar H. The correlation between the spread of COVID-19 infections and weather variables in 30 Chinese provinces and the impact of Chinese government mitigation plans. Eur Rev Med Pharmacol Sci 2020; 24: 4565-4571.
5) Y Xing, H Wang, X Yao, Y Li, J Huang, J Tang, S Zhu, Y Liu, J Xiao. Analysis of factors for disease progression in 61 patients with COVID-19 in Xiaogan, Hubei, China. Eur Rev Med Pharmacol Sci 2020; 24: 12490-12499.
6) Trivedi N, Verma A, Kumar D. Possible treatment and strategies for COVID-19: review and assessment. Eur Rev Med Pharmacol Sci 2020; 24: 12593-12608.
7) Vetrugno G, Laurenti P, Franceschi F, Foti F, D'Ambrosio F, Cicconi M, LA Milia DI, Di Pumpo M, Carini E, Pascucci D, Boccia S, Pastorino R, Damiani G, De-Giorgio F, Oliva A, Nicolotti N, Cambieri A, Ghisellini R, Murri R, Sabatelli G, Musolino M, Gasbarrini A; Gemelli-Against-Covid Group. Gemelli decision tree Algorithm to Predict the need for home monitoring or hospitalization of confirmed and unconfirmed COVID-19 patients (GAP-Covid19): preliminary results from a retrospective cohort study. Eur Rev Med Pharmacol Sci 2021; 25: 2785-2794.
8) Singh V, Poonia C, Kumar S, Dass P, Agarwal P, Bhatnagar V, Raja L. Prediction of COVID-19 corona virus pandemic based on time series data using Support Vector Machine. J Discret Math Sci Cryptogr 2020; 23: 1583-1597.
9) Niazkar R, Niazkar M. Application of artificial neural networks to predict the COVID-19 outbreak. Glob Health Res Policy 2020; 5: 1-11.
10) Al-Rousan N, Al-Najjar H, Alomari O. Assessment of predicting hourly global solar radiation in Jordan based on rules, trees, meta, lazy and function prediction methods. Sustain Energy Technol 2021; 44: 1-14.