A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea

H. AL-NAJJAR, N. AL-ROUSAN

Department of Computer Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, Istanbul, Turkey

Abstract. – OBJECTIVE: Coronavirus COVID-19 further transmitted to several countries globally. The status of the infected cases can be determined basing on the treatment process along with several other factors. This research aims to build a classifier prediction model to predict the status of recovered and death coronavirus CovID-19 patients in South Korea.

MATERIALS AND METHODS: Artificial neural network principle is used to classify the collected data between February 20, 2020 and March 9, 2020. The proposed classifier used different seven variables, namely, country, infection reason, sex, group, confirmation date, birth year, and region. The most effective variables on recovered and fatal cases are analyzed based on the neural network model.

RESULTS: The results found that the proposed predictive classifier efficiently predicted recovered and death cases. Besides, it is found that discovering the infection reason would increase the probability to recover the patient. This indicates that the virus might be controllable based on infection reasons. In addition, the earlier discovery of the disease affords better control and a higher probability of being recovered.

CONCLUSIONS: Our recommendation is to use this model to predict the status of the patients globally.

Key Words:
Epidemiology, Engineering and technology, Infection, South Korea.

Introduction

COVID-19 is firstly reported in China on January 22, 2020; afterward, the disease exported to 162 territories[1]. China, Italy, Iran, Spain, and South Korea are the top five countries that infected from COVID-19[2]. CoVID-19 rapidly transmitted to the nearest countries (i.e., South Korea) and to the farthest countries (i.e., Bolivia, Brazil, Peru, etc.) as well. Thus, understanding and tracing patients' collected information will simplify understanding the main factors and reasons of infections. South Korea in the last five days reported more recovery cases than new infections, as number of recovery cases is increased, the number of confirmed cases started to slow down. The decrease of the daily raised trend was due to the ability of South Korea to control one of the biggest outbreaks outside China[3]. The first death case was reported on February 20, 2020; then, 81 cases were reported by the end of March 17, 2020. The first recovered case was reported on February 7, 2020 and reached 1137 cases by the end of March 17, 2020. No recovered cases showed any reported death cases until March 7, 2020. Most confirmed cases found to have similar symptoms, namely flu, and pneumonia, where other cases did not suffer from the common symptoms[4].

Materials and Methods

This study uses official time series data from the Korea Centers for Disease Control and Prevention (KCDC) for 7869 coronavirus patients in South Korea between 20/01/2020 and 09/03/2020[5]. The dataset contains fifteen variables including patient ID, sex, birth year, country, region, disease, group, infection reason, infection order, infected by, contact number, confirmation date, released date, deceased date, and state. This study adopted seven variables as independent variables including sex, birth year, country, region, group, infection reason, and confirmed date, where dependent variable is one of the following variables, namely death or recovered. The variables are chosen based on the most used variables in several researches. To avoid missing independent and dependent variables, only 659 and 649 patients are employed for recovered and death cases, respectively.

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

To build a classification model, a dataset is divided into training and testing. 70% and 30% of data are selected as training and testing, respectively. The study chooses a neural network to build a classifier with one hidden layer and gradient descent as an optimization algorithm. A neural network is considered as the most efficient prediction model in building a medical classification. To evaluate the efficiency of the selected variables in classifying recovered or death cases, a confusion matrix is used. Moreover, to understand the most effective variables from the selected variables an importance level is applied on independent variables.

Results

The dataset of recovered cases is divided into 466 and 193 for training and testing, respectively, where for death cases, the training and testing are 463 and 186, respectively. The results of predicting death and recovered cases are shown in Table I. The overall accuracy of training process is 95.7%, where 67.5% and 98.4% are accurately classified as not recovered and recovered cases respectively. For testing data, the overall accuracy is 93.8%, where 42.9% and 97.8% are accurately classified as not recovered and recovered cases, respectively. The results revealed that using sex, birth year, country, region, group, infection reason, and confirmed date are efficient to classify the recovered cases. Besides, this indicates that the recovered cases are sensitive to independent variables. For death cases, the overall accuracy of training is 99.6%, where 99.8% and 96.0% are correctly classified as not death and death, respectively. For testing data, the overall classification is 99.5%, where 100% and 85.7% are correctly classified as not death and death, respectively.

The proposed models for death and recovered cases showed the capability of the proposed model in classifying the death and recovered cases based on the studied variables. Therefore, to show the effectiveness of each variable, a variable importance analysis is used as shown in Figure 1. The importance variable analysis of recovered cases showed that the effective variables based on least to most effective are sex, group, country, birth, region, confirmed date, and infection reason. The recovered cases analysis showed that infection reason is the most effective, where sex is the least effective. The results revealed that the infection reason is important to determine the recovered cases; this indicates that different cases might have different CoVID-19 sequences and each infection reason might generate a new sequence. Confirming the infection would help doctors in treating a patient. Birth date is important to determine the possibility of treating, since 87.5% of recovered cases are less than 60 years old. Determining region, country and group would help doctors in determining the CoVID-19 transmission way, tracking other infected patients, and analyzing the CoVID-19 proteins and genomes.

The importance variable analysis of death cases showed that the effective variables based on least to most effective are country, infection reason, sex, group, confirmed date, birth, and region. The death cases analysis showed that the region is the most effective, where country is the least effective. The results revealed that the region is

Table I. Classification results using death and recovered cases as dependent variable.

Dependent variable	Sample		Predicted		Accuracy
			.00	1.00	
Recovered	***Training***	.00	419	7	98.4%
		1.00	13	27	67.5%
		Overall	92.7%	7.3%	95.7%
	Testing	.00	175	4	97.8%
		1.00	8	6	42.9%
		Overall	94.8%	5.2%	93.8%
Death	***Training***	0	437	1	99.8%
		1	1	24	96.0%
		Overall	94.6%	5.4%	99.6%
	Testing	0	184	0	100.0%
		1	1	6	85.7%
		Overall	96.9%	3.1%	99.5%

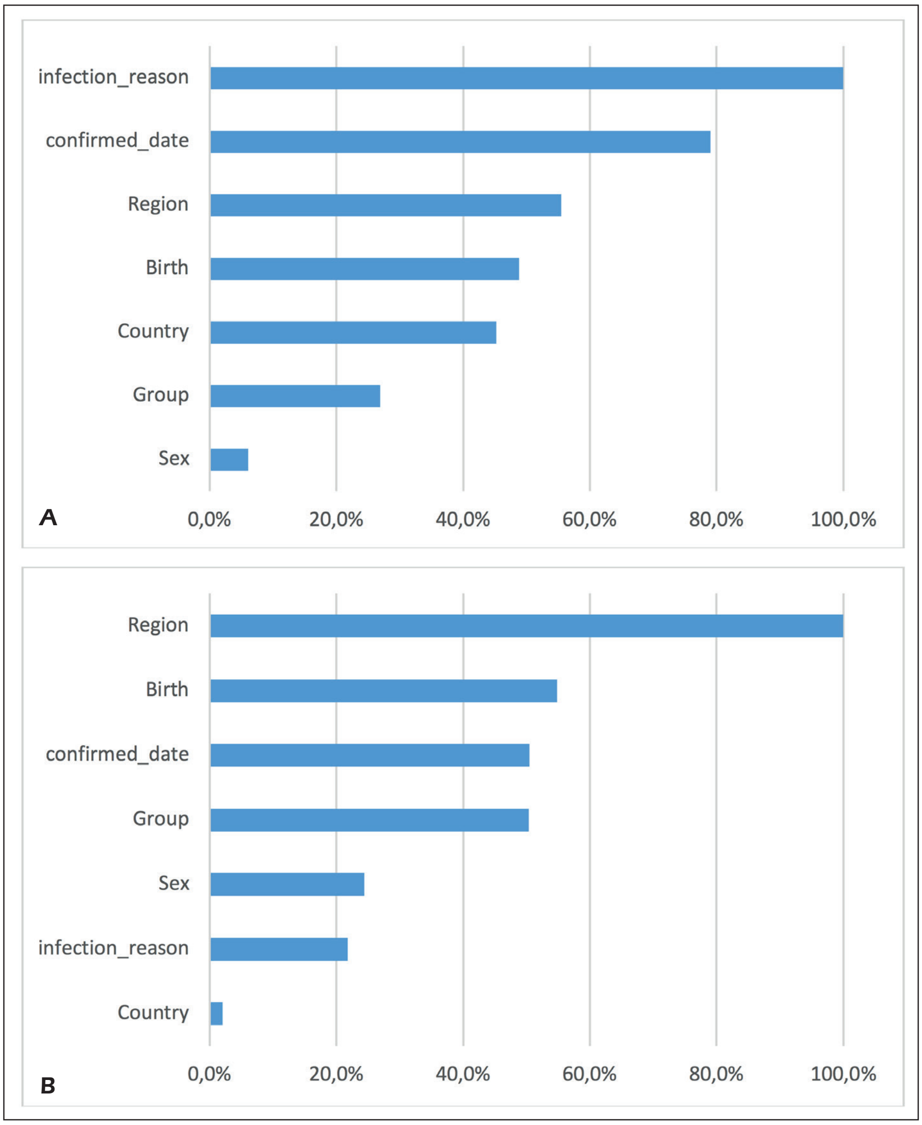

Figure 1. Importance analysis of independent variables to predict (**A**) recovered, and (**B**) death.

important to determine the death cases; this indicates that different regions may have different responses on CoVID-19 (i.e., changing temperature, humidity, and so forth)[6]. While birth year is considered as a second important factor, since many elderlies have many medical problems, which increase the probability to die. Confirmed date is important to fast the process of treatment since late date of confirmation will increase the number of deaths, while group gives information about the place of infection. Group is important because it gives an indicator of the mobility of humans and their contact, while the rest of the variables are less than 30%.

Discussion

It is found that the top three important variables to predict the status of death patients are the infection reason, confirmation date, and region. Where region, birth year, and confirmation date are important for recovered cases. Thus, region and confirmation date are the most effected variables in determining both recovered and death cases. Sex and group are the least important variables to death cases, while infection reason and country are the least important variables to recovered cases. Finally, the results revealed that choosing the most effective categorical variables (non-numerical variables) with numerical variables could enhance the prediction model. Besides, adding variables with the least importance can stabilize a neural network predictor, avoiding overfitting, and improving the prediction output.

Conclusions

This research implemented a classifier that successfully can classify whether the current patients will be in recovered or death groups based on several variables. Understanding the relationship between the status of the patient and other variables (i.e., sex, region, country, and infection reason) is very important. Our classifier could be used for several variables in different countries. Besides, it could provide useful information to stop the spreading of CoVID-19 and to predict the status of each patient separately.

Conflict of Interests

The authors declare that they have no known competing for financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank European Review for Medical and Pharmacological Sciences for allowing the quick publication of this article without a fee.

References

1) Perrella A, Carannante N, Berretta M, Rinaldi M, Maturo N, Rinaldi L. Editorial–Novel Coronavirus 2019 (Sars-CoV2): a global emergency that needs new approaches?. Eur Rev Med Pharmacol Sci 2020; 24: 2162-2164.
2) Meo SA, Alhowikan AM, Al-Khlaiwi T, Meo IM, Halepoto DM, Iqbal M, Usmani AM, Hajjar W, Ahmed N. Novel coronavirus 2019-nCoV: prevalence, biological and clinical characteristics comparison with SARS-CoV and MERS-CoV. Eur Rev Med Pharmacol Sci 2020; 24: 2012-2019.
3) Phelan AL, Katz R, Gostin LO. The novel coronavirus originating in Wuhan, China: challenges for global health governance. JAMA 2020; 323: 709-710.
4) Kannan S, Shaik Syed Ali P, Sheeza A, Hemalatha K. COVID-19 (Novel Coronavirus 2019)–recent trends. Eur Rev Med Pharmacol Sci 2020; 24: 2006-2011.
5) Korea Centers for Disease Control and Prevention (2020). http://ghdx.healthdata.org/organizations/korea-centers-disease-control-and-prevention-kcdc.
6) Al-Rousan N, Al-Najjar H. Nowcasting and forecasting the spreading of Novel Coronavirus 2019-nCoV and its association with weather variables in 30 Chinese provinces: a case study. SSRN. 3537084. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3537084.