Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

© 2022 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220569

Symptom-Based COVID19 Screening Model Combined with Surveillance Information

Dohyung LEE [a], Myeongchan KIM [a], Hyunwoo CHOO [b], Soo-Yong SHIN [b, 1]
[a] *Mobile Doctor Co. Ltd, South Korea*
[b] *Department of Digital Health, SAIHST, Sungkyunkwan University*

Abstract. As the number of cases for COVID-19 continues to grow unprecedentedly, COVID-19 screening is becoming more important. In this study, we trained machine learning models from the Israel COVID-19 dataset and compared models that used surveillance indices of COVID-19 and those that did not. The AUC scores were 0.8478±0.0037 and 0.8062±0.005 with and without surveillance information, respectively, and there was significant improvement when the surveillance information was used.

Keywords. Machine Learning, COVID-19, Symptomatic screening

1. Introduction

As the number of confirmed cases of COVID-19 continues to grow unprecedentedly, effective triaging methods can mitigate the burden on healthcare systems. In previous studies, it has been reported that impressive detecting accuracy can be achieved even with a limited number of features through data publicly reported by the Israeli Ministry of Health [1,2]. However, epidemiological factors mainly influence the risk of being infected by contagious diseases. Thus, the symptom-based model has a drawback in that it does not consider epidemiological factors. In this study, we trained machine learning models with symptom data combined with surveillance information, which is one of the multiple epidemiological factors, and compared models which do not. Hence, our method can enhance the performance of machine learning models.

2. Methods

We utilized symptom data and PCR test results from COVID-19 Dataset by the Israeli Ministry of Health (Israel Dataset) and number of cumulative confirmed COVID-19 cases from Data Repository collected by the Center for Systems Science and Engineering at Johns Hopkins University (CSSE Dataset) from 2020.03.11 to 2021.12.31 [2,3]. Using the number of confirmed cases each day, we generated a national surveillance index and a global surveillance index by min-max normalization from the entire period for Israel and worldwide, respectively. We collated the dataset with 12 features in the following

[1] Soo-Yong Shin, 115, Irwon-ro, Gangnam-gu, Seoul, Republic of Korea; E-mail: sy.shin@skku.edu

list: sex, age (>=60 or <60 years), presence of 5 symptoms, (cough, fever, sore throat, shortness of breath and headache), binary values of 3 test indications ("contact with confirmed cases", "abroad" or "others"), and 2 surveillance indices on the PCR test date (national and global). Out of a total of 7,310,559 entries, we randomly sampled 300,000 entries and split 100,000 entries into training, validation and test datasets. We used a logistic regression model (LR) and 16 layers of 1-dimensional convolutional neural network model with skipped layer (1D ResNet). To see the effect of the surveillance indices on the models, we repeated experiments with and without the surveillance indices (12 features and 10 features, respectively) 5 times for each model. Then we performed a two-sample t-test to compare with and without the surveillance indices.

3. Results

The AUC score of logistic regression with surveillance indices was 0.8374±0.0019, and the AUC score was 0.7828±0.0031 without the indices. The score for 1D ResNet was 0.8478±0.0037 when the surveillance indices were included and 0.8062±0.005 when not included. In the presence of the surveillance indices, both models significantly improved the performance (P<.001). Figure 1. shows the average AUROC of both results.

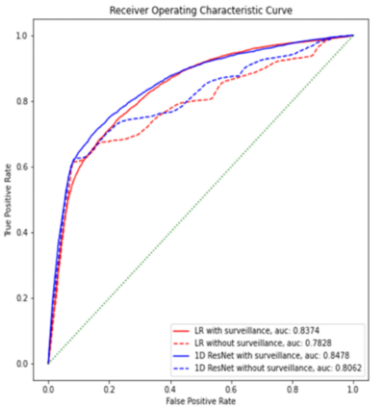

Figure 1.

4. Discussion and Conclusions

Performance of our two models for COVID-19 diagnosis prediction improved when the training data included national and global surveillance index. In the early epidemics, a prior study trained a model with 10 days of Israeli Data, but when this model was tested after more than a year, the performance was not as good written in the study [1]. This may be due to changes in epidemiological characteristics such as variant viruses, and vaccination rates. Since clinical assessment considers epidemiological factors, we also considered them in this study. In addition, we expect to improve performance by adding other epidemiological information such as vaccination rates and local surveillance information.

5. References

[1] Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medicine, 2021;4(1),1–5.
[2] COVID-19-Government Data, 2020, https://data.gov.il/dataset/covid-19
[3] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534