![Research Square]

Development of a Deep Learning System To Reduce The Time Needed for COVID-19 RT-PCR: *Can We Use Deep Learning To Get RT-PCR Test Results of COVID-19 Faster?*

YOONJE LEE

  Hallym University College of Medicine

Yu-Seop KIM

  Hallym University

Da-in Lee

  Hallym University

Seri Jeong (✉ hehebox@naver.com)

  Hallym University College of Medicine

Gu-Hyun Kang

  Hallym University College of Medicine

Yong Soo Jang

  Hallym University College of Medicine

Wonhee Kim

  Hallym University College of Medicine

Hyun Young Choi

  Hallym University College of Medicine

Jae Guk Kim

  Hallym University College of Medicine

Sang-hoon Choi

  Yonsei University

Research Article

Abstract

Reducing the time to diagnose COVID-19 helps to manage insufficient isolation-bed resources and adequately accommodate critically ill patients in clinical fields. There is currently no alternative method to RT-PCR, which requires 40 cycles to diagnose COVID-19. We proposed a deep learning (DL) model to improve the speed of COVID-19 RT-PCR diagnosis. We developed and tested a DL model using the long-short term memory method with a dataset of fluorescence values measured in each cycle of 5,810 RT-PCR tests. Among the DL models developed here, the diagnostic performance of the 21st model showed an area under the receiver operating characteristic (AUROC), sensitivity, and specificity of 84.55%, 93.33%, and 75.72%, respectively. The diagnostic performance of the 24th model showed an AUROC sensitivity, and specificity of 91.27%, 90.00%, and 92.54%, respectively.

Introduction

Last year, 2020, was a time when all humanity grappled with new changes due to the coronavirus disease 2019 (COVID-19) pandemic. COVID-19 has presented a serious threat due to despite being caused by infectious agents that represent a miniscule mass even when combined. Even in this time of confusion, humanity made every effort to survive, overcome the crisis and find order.

From a medical perspective, the COVID-19 pandemic has required us to be fast and accurate in four kinds of approaches: diagnosis, isolation, treatment, and tracking. Among these, a fast diagnosis is of paramount importance because the remaining 3 approaches can proceed quickly and accurately only when a fast diagnosis precedes them. This requirement is because the response must be able to quickly block the spread of infection by rapidly applying the remaining 3 approaches upon rapid diagnosis.

For the diagnosis of this infectious disease, the real-time reverse transcriptase polymerase chain reaction (RT-PCR) test is most widely used as a reference test. Unfortunately, despite the importance of rapid diagnosis, this test can take up to approximately 6 hours from sampling and may require consecutive tests to discriminate false negative or positive results[1].

Various efforts have been made to quickly and accurately diagnose COVID-19 with the help of machine learning or artificial intelligence (AI) using information such as symptoms, chest X-ray (CXR) findings, computed tomography (CT) findings, and routine laboratory blood test results[2-7]. Clinically useful results have been reported; however, they cannot completely replace the RT-PCR test[2-7]. In addition, rapid diagnostic kits for detecting severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), such as loop-mediated isothermal amplification assays[8] and immunoassays utilizing antigen-antibody responses[9,10], have also been developed. However, their relative performances compared to those of conventional RT-PCR are limited, and the available infrastructure for sufficient application to suspected patients is not available.

Unlike the approaches of previous studies, we tried to investigate whether the time taken for RT-PCR diagnosis can be reduced through a deep learning (DL) model developed in this study. We intended to develop a DL model using raw data of fluorescence values in every cycle of RT-PCR that can predict the results before completion of the RT-PCR test and tested this model.

Methods

Study participants

We enrolled patients who visited a specialized outpatient department for COVID-19 triage or emergency department to identify COVID-19 between 23 November 2020 and 19 January 2021. The raw data of RT-PCR curves determined to detect SARS-CoV-2 during this period were collected.

Of the 5,810 patients' data included in the study, 181 had positive RT-PCR results, while 5,629 had negative results. These data were divided into two datasets for training and testing (Figure 1).

This study was approved by the Institutional Review Committee (HKS 2020-07-007) of Hallym University Gangnam Sacred Heart Hospital in Korea with a waived consent because the subject's data were anonymized. And this study was conducted in accordance with the STARD guideline and regulation as a study related to the diagnostic accuracy of COVID-19 RT-PCR.

Materials

The reagent for RT-PCR assay used in this study was the STANDARD M nCoV Real-Time Detection kit (SD Biosensor, Gyeonggi, South Korea), and the analyzer was Bio-Rad CFX96 (Bio-Rad Laboratories, Inc., Hercules, CA, USA).

Development of a DL model

The RT-PCR results (positive or negative) were used as the output variable to train the models. A total of 40 models were developed and validated, from the model trained with the fluorescence value of the RT-PCR first cycle to the model trained from the fluorescence value of the entire 40 RT-PCR cycles.

For example, the first model was trained with the fluorescence value of the first RT-PCR cycle, and the second model was trained with the fluorescence values from the first to second RT-PCR cycles. In the same way, the 39th model was trained with the fluorescence values from the first to the 39th RT-PCR cycle, and the 40th model was trained with the fluorescence values from the first to the 40th RT-PCR cycle.

Since the fluorescence values derived in the RT-PCR process have the characteristics of time series data, we developed a total of 40 DL models using LSTM (Figure 2). All deep learning analyses were performed using Python.

Training and test datasets

The results of the RT-PCR virology test were used as the reference to train the models. The data for training were composed of curves of RT-PCR results of 91 positive cases and 2,814 negative cases. The data of 90 positive and 2,815 negative cases were used for testing (Figure 1).

Outcomes

Primary outcomes were the sensitivity, specificity and AUROC values of each model.

Secondary outcomes suggested an optimal model using PPV, NPV and accuracy according to the prevalence of each model for several countries: the United States, Italy, and South Korea. The prevalence data for each country were referenced from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE)" at Johns Hopkins University[11]. The prevalence was based on values measured for each country between June and July 2021. In a triangular-shaped radar chart using the PPV, NPV, and accuracy values affected by the prevalence to visualize diagnostic performance, each model was compared by calculating the ratio of the area of the triangle covered by each model to the total triangle area of the radar chart (Fig. 4~6).

Statistical analysis

All statistical analyses were performed using SPSS software V.26.0 (IBM, SPSS Inc., Chicago, IL, United States). Sensitivity (the proportion of true positives) and specificity (the proportion of true negatives) were calculated in comparison with the positivity or negativity of RT-PCR results. We calculated false positive and false negative rates using the confusion matrix and calculated the PPVs and NPVs of each model using the COVID-19 prevalence for several countries: the United States, Italy and South Korea.

Results

Diagnostic performance of DL models

Table 1 shows the diagnostic performance of the data from the 20th model trained with raw data (1 to 20 cycles) to the 36th model trained with raw data (1 to 36 cycles). The sensitivity was the highest at 100% (95% CI; 95.98–100.0%) in the 33rd model and the 34th model, and the specificity was the highest at 96.77% (95% CI; 96.05–97.39%) in the 25th model. The AUROC value was the highest at 97.00% in the 33rd model, which had the highest sensitivity. The AUROC value of the 25th model, which had the highest specificity, was 88.38%. The model with the lowest false positive rate (FPR) was the 25th model at 3.23%, and the model with the lowest false negative rate (FNR) was the 33rd model at 0%.

Table 1
Diagnostic performance of deep learning models

Model	Sensitivity, % (95% CI; %)	Specificity, %, (95% CI; %)	AUROC, %	False Positive, %	False Negative, %
20	76.53 (74.00 to 90.36)	68.56 (66.81 to 70.27)	75.95	31.44	16.67
21	93.33 (86.05 to 97.51)	75.72 (74.15 to 77.35)	84.55	24.23	6.67
22	66.67 (55.95 to 76.26)	89.38 (88.18 to 90.49)	78.02	10.62	33.33
23	83.33 (74.00 to 90.36)	92.40 (91.36 to 93.35)	87.87	7.60	16.67
24	90.00 (81.86 to 95.32)	92.54 (91.51 to 93.48)	91.27	7.46	10.00
25	80.00 (70.25 to 87.69)	96.77 (96.05 to 97.39)	88.38	3.23	20.00
26	92.22 (84.63 to 96.82)	91.69 (90.61 to 92.68)	91.95	8.31	7.78
27	93.33 (86.05 to 97.51)	92.58 (91.54 to 93.52)	92.95	7.42	6.67
28	96.67 (90.57 to 99.31)	93.25 (92.26 to 94.15)	94.96	6.75	3.33
29	94.44 (87.51 to 98.17)	85.04 (83.67 to 86.34)	89.74	14.96	5.56
30	96.67 (90.57 to 99.31)	91.08 (89.97 to 92.11)	93.88	8.92	3.33
31	97.78 (92.20 to 99.73)	90.55 (89.41 to 91.61)	94.16	9.45	2.22
32	96.67 (90.57 to 99.31)	92.68 (91.66 to 93.62)	94.67	7.32	3.33
33	100.00 (95.98 to 100.0)	94.00 (93.05 to 94.85)	97.00	6.00	0
34	100.00 (95.98 to 100.0)	93.32 (92.34 to 94.22)	96.66	6.68	0
35	97.78 (92.20 to 99.73)	95.52 (94.69 to 96.26)	96.65	4.48	2.22
36	98.89 (93.96 to 99.97)	93.07 (92.07 to 93.98)	95.98	6.93	1.11
CI: Confidence interval; AUROC: Area under the receiver operating characteristic					

Comparison of the diagnostic performance of each model and RT-PCR

The test results of 40 DL models developed through long-short term memory (LSTM) are shown in Figure 3. The sensitivity of each model was 93.33% (95% CI; 86.05−97.51%) in the model constructed using raw data up to 21 cycles, considering the sensitivity reference value of 89% of RT-PCR. However, the 21st model had a low specificity of 75.72% (95% CI; 74.15−77.35%). The twenty-fourth model showed sensitivity, specificity and AUROC values of 90% (95% CI; 81.86−95.32%), 92.54% (95% CI; 91.51−93.48%) and 91.27%, respectively (Figure 3).

Effects of prevalence on screening performance each model: United States, Italy, South Korea

In the United States, showing a prevalence of 10.06%, the model with the highest positive predictive value (PPV) was the 25th model at 73.33% (95% CI; 68.67−77.53%), and the model with the highest negative predictive value (NPV) was the 33rd model at 100% (95% CI; N/A). Accuracy was the highest at 95.75% (95% CI; 94.95 to 96.45) in the 35th model. The NPV of the 25th model with the highest PPV was 97.76% (95% CI; 96.64−98.50%), and the accuracy was 95.09% (95% CI; 94.24−95.85%). The PPV of the 33rd model with the highest NPV was 64.92% (95% CI; 61.52−68.17%), and the accuracy was 94.60% (95% CI; 93.71−95.39%) (Table 2).

Table 2
Effects of prevalence on screening performance each model: United States (Prevalence 10.06%)

Model	PPV, % (95% CI; %)	NPV, % (95% CI; %)	Accuracy, % (95% CI, %)
20	21.29 (20.92 to 24.69)	96.34 (95.89 to 98.33)	69.36 (68.34 to 71.70)
21	29.93 (28.21 to 31.80)	99.03 (97.93 to 99.55)	77.48 (75.97 to 79.04)
22	41.09 (36.78 to 45.53)	96.02 (94.74 to 97.00)	87.11 (85.83 to 88.31)
23	54.91 (50.97 to 58.80)	98.04 (96.92 to 98.75)	91.49 (90.42 to 92.48)
24	57.27 (53.64 to 60.83)	98.81 (97.82 to 99.36)	92.29 (91.26 to 93.23)
25	73.33 (68.67 to 77.53)	97.76 (96.64 to 98.50)	95.09 (94.24 to 95.85)
26	55.21 (51.81 to 58.56)	99.07 (98.12 to 99.54)	91.74 (90.68 to 92.72)
27	58.28 (54.80 to 61.68)	99.21 (98.30 to 99.63)	92.65 (91.64 to 93.57)
28	61.41 (57.98 to 64.73)	99.60 (98.81 to 99.87)	93.59 (92.64 to 94.46)
29	41.23 (38.80 to 43.71)	99.28 (98.33 to 99.69)	85.98 (84.67 to 87.23)
30	54.64 (51.55 to 57.70)	99.60 (98.78 to 99.87)	91.64 (90.58 to 92.62)
31	53.48 (50.53 to 56.42)	99.73 (98.94 to 99.93)	91.27 (90.19 to 92.27)
32	59.48 (56.14 to 62.73)	99.60 (98.80 to 99.87)	93.08 (92.10 to 93.98)
33	64.92 (61.52 to 68.17)	100.00 (N/A)	94.60 (93.71 to 95.39)
34	62.46 (59.17 to 65.64)	100.00 (N/A)	93.99 (93.06 to 94.83)
35	70.82 (67.11 to 74.27)	99.74 (98.99 to 99.93)	95.75 (94.95 to 96.45)
36	61.33 (58.03 to 64.53)	99.87 (99.08 to 99.98)	93.65 (92.71 to 94.51)
CI: Confidence interval; PPV: Positive predictive value; NPV: Negative predictive value; N/A: Not applicable			

In Italy, which showed a prevalence of 6.98%, the model with the highest PPV was the 25th model at 65.02% (98% CI; 59.68−69.97%), and the model with the highest NPV was the 33rd model at 100% (95% CI; N/A). Accuracy was the highest at 95.68% (95% CI; 94.88−96.39%) in the 35th model. The NPV of the 25th model with the highest PPV was 98.47% (95% CI; 97.71−98.98%), and the accuracy was 95.60% (95% CI; 94.79−96.31%). The PPV of the 33rd model with the highest NPV was 55.57% (95% CI; 51.92−59.13%), and the accuracy was 94.42% (95% CI; 93.52−95.22%) (Table 3).

Table 3
Effects of prevalence on screening performance each model: Italy (Prevalence 6.98%)

Model	PPV, % (95% CI; %)	NPV, % (95% CI; %)	Accuracy, % (95% CI; %)
20	16.60 (15.16 to 18.13)	98.21 (97.18 to 98.86)	69.59 (67.88 to 71.26)
21	22.44 (20.97 to 23.95)	99.34 (98.59 to 99.70)	77.00 (75.42 to 78.52)
22	32.03 (28.21 to 36.08)	97.28 (96.39 to 97.95)	87.79 (86.55 to 88.96)
23	45.15 (41.24 to 49.08)	98.66 (97.90 to 99.15)	91.76 (90.71 to 92.74)
24	47.53 (43.86 to 51.19)	99.20 (98.52 to 99.57)	92.36 (91.34 to 93.30)
25	65.02 (59.68 to 69.97)	98.47 (97.71 to 98.98)	95.60 (94.79 to 96.31)
26	45.45 (42.07 to 48.83)	99.37 (98.72 to 99.69)	91.72 (90.66 to 92.70)
27	48.56 (45.02 to 52.08)	99.46 (98.84 to 99.75)	92.63 (91.62 to 93.55)
28	51.82 (48.24 to 55.34)	99.73 (99.19 to 99.91)	93.49 (92.53 to 94.36)
29	32.17 (29.98 to 34.40)	99.51 (98.86 to 99.79)	85.70 (84.37 to 86.95)
30	44.88 (41.81 to 47.94)	99.73 (99.17 to 99.91)	91.74 (90.40 to 92.46)
31	43.73 (40.82 to 46.64)	99.82 (99.28 to 99.95)	91.06 (89.96 to 92.07)
32	49.80 (46.36 to 53.20)	99.73 (99.19 to 99.91)	92.96 (91.97 to 93.86)
33	55.57 (51.92 to 59.13)	100.00 (N/A)	94.42 (93.52 to 95.22)
34	52.93 (49.46 to 56.33)	100.00 (N/A)	93.79 (92.85 to 94.64)
35	61.13 (57.95 to 66.10)	99.83 (99.32 to 99.96)	95.68 (94.88 to 96.39)
36	51.72 (48.29 to 55.13)	99.91 (99.37 to 99.99)	93.48 (92.52 to 94.35)
CI: Confidence interval; PPV: Positive predictive value; NPV: Negative predictive value; N/A: Not applicable			

In South Korea, which showed a prevalence of 0.27%, the model with the highest PPV was the 25th model at 6.43% (95% CI; 5.07−7.76%), and the model with the highest NPV was the 33rd model at 100% (95% CI; N/A). Accuracy was the highest at 96.72% (95% CI; 96.01−97.34%) in the 25th model. The NPV of the 25th model with the highest PPV was 99.94% (95% CI; 99.92−99.96%). The PPV of the 33rd model with the highest NPV was 4.42% (95% CI; 3.75−4.96%), and the accuracy was 94.01% (95% CI; 93.09−94.85%) (Table 4).

Table 4
Effects of prevalence on screening performance each model: South Korea (Prevalence 0.27%)

Model	PPV, % (95% CI; %)	NPV, % (95% CI; %)	Accuracy, % (95% CI; %)
20	0.67 (0.64 to 0.79)	99.91 (99.90 to 99.96)	68.58 (66.88 to 70.29)
21	1.06 (0.95 to 1.12)	99.98 (99.95 to 99.99)	75.77 (74.22 to 77.37)
22	1.71 (1.40 to 2.00)	99.90 (99.86 to 99.92)	89.32 (88.14 to 90.42)
23	2.95 (2.47 to 3.36)	99.95 (99.92 to 99.97)	92.37 (91.35 to 93.31)
24	3.24 (2.74 to 3.65)	99.97 (99.95 to 99.98)	92.53 (91.52 to 93.46)
25	6.43 (5.07 to 7.76)	99.94 (99.92 to 99.96)	96.72 (96.01 to 97.34)
26	2.99 (2.55 to 3.33)	99.98 (99.95 to 99.99)	91.69 (90.63 to 92.67)
27	3.37 (2.87 to 3.77)	99.98 (99.96 to 99.99)	92.58 (91.56 to 93.50)
28	3.83 (3.25 to 4.28)	99.99 (99.97 to 100.00)	93.26 (92.29 to 94.14)
29	1.72 (1.52 to 1.86)	99.98 (99.96 to 99.99)	85.07 (83.72 to 86.35)
30	2.92 (2.53 to 3.22)	99.99 (99.97 to 100.00)	91.10 (90.00 to 92.11)
31	2.79 (2.43 to 3.06)	99.99 (99.97 to 100.00)	90.57 (89.45 to 91.61)
32	3.54 (3.02 to 3.94)	99.99 (99.97 to 100.00)	92.69 (91.69 to 93.61)
33	4.42 (3.75 to 4.96)	100.00 (N/A)	94.01 (93.09 to 94.85)
34	3.99 (3.41 to 4.45)	100.00 (N/A)	93.34 (92.37 to 94.22)
35	5.72 (4.74 to 6.57)	99.99 (99.98 to 100.00)	95.53 (94.71 to 96.25)
36	3.72 (3.26 to 4.24)	100 (99.98 to 100.00)	93.09 (92.10 to 93.98)

CI: Confidence interval; PPV: Positive predictive value; NPV: Negative predictive value; N/A: Not applicable

Comparison of each model considering the effects of prevalence on screening performance: United States, Italy, South Korea

In the United States, which showed a prevalence of 10.06%, the model with the lowest PPV was the 1st model, at 8.52% (95% CI; 4.08−16.92%), and the highest PPV was that of the 25th model at 73.33% (95% CI; 68.67−77.53%). The model with the lowest NPV was the 1st model, at 89.87% (95% CI; 89.35−90.37%), and the highest PPV was that of the 33rd model at 100% (95% CI; N/A). Accuracy was the lowest at 48.48% (95% CI; 46.65−50.32%) in the 3rd model and the highest at 95.75% (95% CI; 94.95−96.45%) in the 35th model. Considering the ratio of area occupied by the radar chart, the 25th model out of the 40 models was the model with the largest proportion of area occupied by the radar chart at 78.13% (95% CI; 74.05−79.96%) (Figure 4).

In Italy, which showed a prevalence of 6.98%, the model with the lowest PPV was the 1st model, at 5.92% (95% CI; 2.79−12.09%), and the highest PPV was that of the 25th model at 65.02% (95% CI; 59.68−69.97%). The model with the lowest NPV was the 1st model, at 92.92% (95% CI; 92.55−93.29%), and the highest NPV was that of the 33rd model at 100% (95% CI; N/A). Accuracy was the lowest at 46.94% (95% CI; 45.11−48.77%) in the 3rd model and the highest at 95.68% (95% CI; 94.88−96.39%) in the 35th model. Considering the ratio of area occupied by the radar chart, the 25th model out of the 40 models was the model with the largest proportion of area occupied by the radar chart at 73.44% (95% CI; 69.17−77.32%) (Figure 5).

In South Korea, which showed a prevalence of 0.27%, the model with the lowest PPV was the 1st model, at 0.23% (95% CI; 0.10−0.49%), and the highest PPV was that of the 25th model at 6.43% (95% CI; 5.07−7.76%). The model with the lowest NPV was the 1st model, at 99.72% (95% CI; 99.71−99.74%), and the highest NPV was that of the 33rd model at 100% (95% CI; N/A). Accuracy was the lowest at 43.52% (95% CI; 41.70−45.34%) in the 3rd model and the highest at 96.72% (95% CI; 96.01−97.34%) in the 25th model. Considering the ratio of area occupied by the radar chart, the 25th model out of the 40 models was the model with the largest proportion of area occupied by the radar chart at 36.44% (95% CI; 35.29−37.54%) (Figure 6).

Discussion

In this study, we developed a total of 40 DL models to reduce the time required for the diagnosis of COVID-19 using RT-PCR as much as possible and compared the diagnostic and screening performance of each model.

In a previous meta-analysis, Kim et al.[12] determined that the the pooled sensitivity of RT-PCR was 89%, and the PPVs and NPVs, affected by the prevalence, were 47.3−98.3% and 93.4−99.9%, respectively. We used the diagnostic performance of the RT-PCR test investigated by Kim et al to compare the performance of each model obtained in this study.

Considering pooled RT-PCR sensitivity of 89% as a sensitivity reference value[12], the sensitivity of the 21st model exceeded this standard and showed 93.33% (95% CI; 86.05−97.51%). In addition, considering the approximate trend of diagnostic performance of all models, the 24th model with a sensitivity of 90% (95% CI; 81.86−95.32%) showed a tendency to exceed the sensitivity reference value. In view of these results, using a Ct value is 36 rather than the time taken by 40 cycles for RT-PCR diagnosis, it can be inferred that meaningful shortening may be possible through the development of this DL model.

Furthermore, the sensitivity reference value also exceeded or showed a similar level from the 3rd model to the 9th model and in the 11th, 16th, and 18th models (Supplementary Table 1). However, the specificities of these models were generally lower than 80%, so it was difficult to judge whether the model was appropriate based on the diagnostic performance.

In the case of the PPV in this study, in the United States, where the prevalence was 10.06%, the 25th model showed the highest PPV at 73.33%. Similarly, in Italy with a prevalence of 6.98% and South Korea with a prevalence of 0.27%, the PPV was highest in the same model as that in the United States at 65.02% and 6.43%, respectively. However, according to the study results of Kim et al.[12], in the United States with a prevalence of 17.7% in March-April 2020, Germany with a prevalence of 5.7%, and Taiwan with a prevalence of 1%, the PPVs of RT-PCR itself were 95%, 84.3% and 47.3%, respectively. Although the prevalence did not match between the two studies and the timing at which the prevalence was measured was different, considering the range of prevalence levels, it can be inferred that the positive screening performance of the model developed in this study is somewhat inferior to that of RT-PCR.

On the other hand, in the case of negative screening performance, which is affected by the prevalence, in the United States, where the prevalence is 10.06%, the 20th model already showed a NPV of 96.34% (95% CI; 95.89−98.33%), and in Italy (prevalence 6.98%) and South Korea (prevalence 0.27%), the NPV was 98.21% (95% CI; 97.18−98.86%) and 99.21% (95% CI; 99.90−99.96%) in the same model, respectively.

According to the research results of Kim et al.[12], the PPV and NPV of RT-PCR showed a distribution of 47.3−98.3% and 93.4−99.9%, respectively, according to the national prevalence (prevalence range of 1−39% from March ~ April 2020). Considering the screening performance of RT-PCR itself, the negative screening performance of the models developed in this study may be considered at a similar level compared to that of RT-PCR.

In this study, we made a radar chart for each model using PPV, NPV and accuracy, which were affected by prevalence, representing screening performance. Then, the screening performance of each model was expressed by expressing the ratio of the area covered by each model to the total area of the radar chart as a percentage, and the area ratio of each model was made into a radar chart. Through this chart, we could confirm that the model with the largest area ratio was the 25th model when considering the PPV, NPV and accuracy. We hypothesize that it would be reasonable to present the 25th model as a model with minimal bias in negative screening performance, positive screening performance and accuracy based on these results.

To the best of our knowledge, no study has reduced the time required to diagnose based on RT-PCR by developing a model trained with raw RT-PCR data and confirming its diagnostic performance. Although there was a study that used RT-PCR curves to build an AI model such as a convolutional neural network (CNN) to reduce false-positive diagnoses, the study was not related to shortening the time for diagnosis and used graph images, differentiating it from our study[13]. In addition, a recently published AI- and DL-related COVID-19 diagnostic study was about a model trained on CT images or CXR images using various CNN methods. Other studies on the diagnosis of COVID-19 have been about a model trained with blood test results or clinical information. First, looking at the studies that reported the performance of models trained based on CNNs using chest CT images, the sensitivity ranged from 77−90%, the specificity ranged from 68−96.6%, and the AUROC ranged from 0.85 to 0.97[1~3, 14~19]. Second, in studies that reported the performance of models trained on CNNs using chest CXR images, the sensitivity ranged from 78−97%,

the specificity ranged from 72.6–99.17%, and the AUROC ranged from 0.77 to 0.92[4~7, 20~22]. Third, there have been studies evaluating the diagnostic performance of COVID-19 using models trained with blood tests or clinical information. In these studies, the sensitivity ranged from 66–93%, the specificity ranged from 64–97.9%, and the AUROC ranged from 0.86 to 0.979[23~25].

What is needed in the clinical field is to increase the efficiency of hospital bed resource management through rapid isolation, rapid diagnosis, and rapid and safe release from isolation. From that perspective, the above studies suggest that COVID-19 diagnosis may be possible through the application of AI. However, due to the imbalance and bias of the data selected for use in training, we question whether this approach can be safely used in clinical settings for the diagnosis of COVID-19. On these issues, Laghi A agrees that efforts to diagnose COVID-19 through AI models are necessary. However, he said that it seems very risky to trust the diagnostic performance of the AI model presented in these studies and use it in clinical settings because imaging tests such as CXR or chest CT at the early stage of COVID-19 infection can show normal findings[26].

The model developed in this study is not a model trained from imaging tests such as CXR or chest CT, blood test results, or clinical information, as in previous studies. In this study, a model trained with LSTM was developed as a DL method applied to time series data training using raw data from 1 to 40 cycles of RT-PCR. There is potential for early diagnosis via RT-PCR using the DL model developed in this study. In this study, the sensitivity of the 21st model had already started to exceed the sensitivity reference value, and the sensitivity and specificity of the 24th model had exceeded 90%. Considering the time it takes to diagnose after 40 cycles of RT-PCR, the diagnostic performance of the model developed in this study shows the possibility of reducing the time taken for RT-PCR diagnosis by almost half. In addition, the model developed in this study showed that the PPV had somewhat lower positive screening performance than RT-PCR; however, the NPV showed negative screening performance similar to that of RT-PCR. If various information, such as the patient's clinical characteristics, blood test results, and imaging information, such as CXR or chest CT results, are combined with this DL model, it can be assumed that the diagnostic performance for early diagnosis will be more sophisticated. We can infer that these efforts have the potential to contribute to improving the efficiency of in-hospital bed resource management for patients with fever or screening symptoms.

There are several limitations in this study. First, 181 positive cases and 5,629 negative cases used for training constituted too few positive cases compared to negative cases. This data bias can affect the diagnostic performance of the developed DL models, and in the end, it is difficult to use the DL model universally. However, through this study, we were able to confirm that the diagnostic performance was not significantly impaired without performing all 40 cycles of PCR. Second, other than LSTM, other DL methods that can be trained using time series data were not applied. As a result, it is not known whether LSTM is the best method because comparative analysis with models that can be developed through other DL methods has not been performed. Nevertheless, LSTM is a recurrent neural network (RNN)-based method that was first selected and used in this study because it is a method created to solve the vanishing gradient problem of existing RNNs[27]. Of course, it is necessary to collect additional data in a

follow-up study and perform comparative analysis with DL methods applied to time series data. Third, the method of presenting the screening performance of the model as the ratio of the area of the radar chart is not a general method. The area of the triangle is calculated assuming the PPV, NPV, and accuracy as a weight of 1:1:1. Therefore, if this weight is set differently, that is, if the three weights are set differently according to need (such as accuracy being more important, etc.), the calculated area and the ratio may be different. Nevertheless, as the PPV, NPV and accuracy all have higher values, it is natural that the screening power increases. We thought the ratio of the area of the radar chart did not perfectly reflect the screening power of the DL model; however, it would help to explain the approximate trend.

Conclusion

Through the test results of the DL models developed in this study, we confirmed the possibility of shortening the diagnosis time of RT-PCR without impairing its diagnostic performance. The reduction in time is expected to be of great help in managing insufficient bed resources in the clinical field.

Declarations

Availability of data and materials

All data generated or analysed during this study are included in this published article (and its supplementary information files).

Competing interests

Seri Jeong, as the corresponding author of this study, declares that there are no conflicts of interest related to this paper. The other authors declare that there are no conflicts of interest associated with this study.

Authors' contributions

Y Lee and YS Kim were involved in the study design, management, data collection, result interpretation and cowriting of the manuscript. Y Lee and YS Kim contributed equally to this study. GH Kang, HY Choi, JG Kim, YS Jang, and W Kim were involved in interpretation of the data and critical revision of the paper for important intellectual content. D Lee and S Choi contributed to the development and analysis of DL models using Python. As the corresponding author, S Jeong was involved in this study concept and design, critical revision of the paper, and final approval of the version to be published.

Acknowledgments

References

1. Mei, X. *et al.* Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nature Medicine*, **26** (8), 1224–1228 (2020 Aug).

2. Jin, C. *et al.* Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. *Nature Communications*, **11**, 1 (2020).

3. Javor, D. *et al.* Deep learning analysis provides accurate COVID-19 diagnosis on chest computed tomography.European Journal of Radiology(2020)133

4. Fontanellaz, M. *et al.* A Deep-Learning Diagnostic Support System for the Detection of COVID-19 Using Chest Radiographs: A Multireader Validation Study. Investigative radiology(2020) 30 Nov 2020

5. Wang, D., Mo, J., Zhou, G., Xu, L. & Liu, Y. An efficient mixture of deep and machine learning models for COVID-19 diagnosis in chest X-ray images. *PLoS ONE*, **15**, 11 (2020).

6. Carlile, M. *et al.* Deployment of Artificial Intelligence for Radiographic Diagnosis of COVID-19 Pneumonia in the Emergency Department. *J Am Coll Emerg Physicians Open*, **1** (6), 1459–1464 (2020 Dec).

7. Zhang, R. *et al. Diagnosis of Coronavirus Disease 2019 Pneumonia by Using Chest Radiography* (Value of Artificial Intelligence, Radiology, 2020).

8. Chaouch, M. Loop-mediated isothermal amplification (LAMP): An effective molecular point-of-care technique for the rapid diagnosis of coronavirus SARS-CoV-2. *Rev Med Virol*, e2115 https://doi.org/10.1002/rmv.2215 (2020).

9. Riccò, M. *et al.* Point-of-care Diagnostic Tests for Detecting SARS-CoV-2 Antibodies: A Systematic Review and Meta-Analysis of Real-World Data. *J. Clin. Med*, **9** (5), 1515 https://doi.org/10.3390/jcm9051515 (2020).

10. Hayera, J., Kasapicb, D. & Zemmrich, C. Real-world clinical performance of commercial SARS-CoV-2 rapid antigen tests in suspected COVID-19: A systematic meta-analysis of available data as of November 20, 2020. *International Journal of Infectious Diseases*, **108**, 592–602 (2021).

11. "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE)" at the Johns Hopkins University". https://github.com/CSSEGISandData/COVID-19

12. Kim, H., Hong, H. & Yoon, S. H. Diagnostic Performance of CT and Reverse Transcriptase Polymerase Chain Reaction for Coronavirus Disease 2019: A Meta-Analysis. Radiology Vol 296 No 3 Sep 2020 E145-155

13. Alouani, D. J., Rajapaksha, R. R. P., Jani, M., Rhoads, D. D. & Sadri, N. Specificity of SARS-CoV-2 Real-Time CPR Improved by Deep Learning Analysis. *Journal of Clinical Microbiology. June*, **59** (6), e02959–20 (2021).

14. Serte, S. & Demirel, H. Deep learning for diagnosis of COVID-19 using 3D CT scans. *Computers in Biology & Medicine*, **132**, 104306 (2021).

15. Yousefzadeh, M. *et al.* ai-corona: Radiologist-assistant deep learning framework for COVID-19 diagnosis in chest CT scans. *PLoS ONE*, **16** (5), e0250952 (2021).

16. Shah, V. *et al.* Diagnosis of COVID-19 using CT scan images and deep learning techniques. *Emerg. Radiol*, **28** (3), 497–505 (2021 Jun).

17. Wu, Z. *et al.* Texture feature-based machine learning classifier could assist in the diagnosis of COVID-19. *European Journal of Radiology*, **137**, 109602 (2021 Apr).

18. Wang, S. *et al.* A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *European Respiratory Journal*, **56** (2), 08 (2020).

19. Li, L. *et al.* Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy.Radiology.296 (2): E65-E71, 2020 08.

20. Wang, G. *et al.* A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nature Biomedical Engineering*, **5** (6), 509–521 (2021 06).

21. Khuzani, A. Z., Heidari, M. & Shariati, S. A. COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images. *Sci. Rep*, **11** (1), 9887 (2021).

22. Castiglioni, I. *et al.* Machine learning applied on chest x-ray can aid in the diagnosis of COVID-19: a first experience from Lombardy, Italy.European Radiology Experimental. 5 (1): 7, 2021 02 02.

23. Alves, M. A. *et al.* Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. *Computers in Biology & Medicine*, **132**, 104335 (2021 05).

24. Kukar, M. *et al.* COVID-19 diagnosis by routine blood tests using machine learning.Scientific Reports. 11 (1):10738, 2021 05 24.

25. Goodman-Meza, D. *et al.* A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. *PLoS ONE*, **15** (9), e0239474 (2020).

26. Laghi, A. Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence. *Lancet Digital Health*, **2**, 5 (2020).

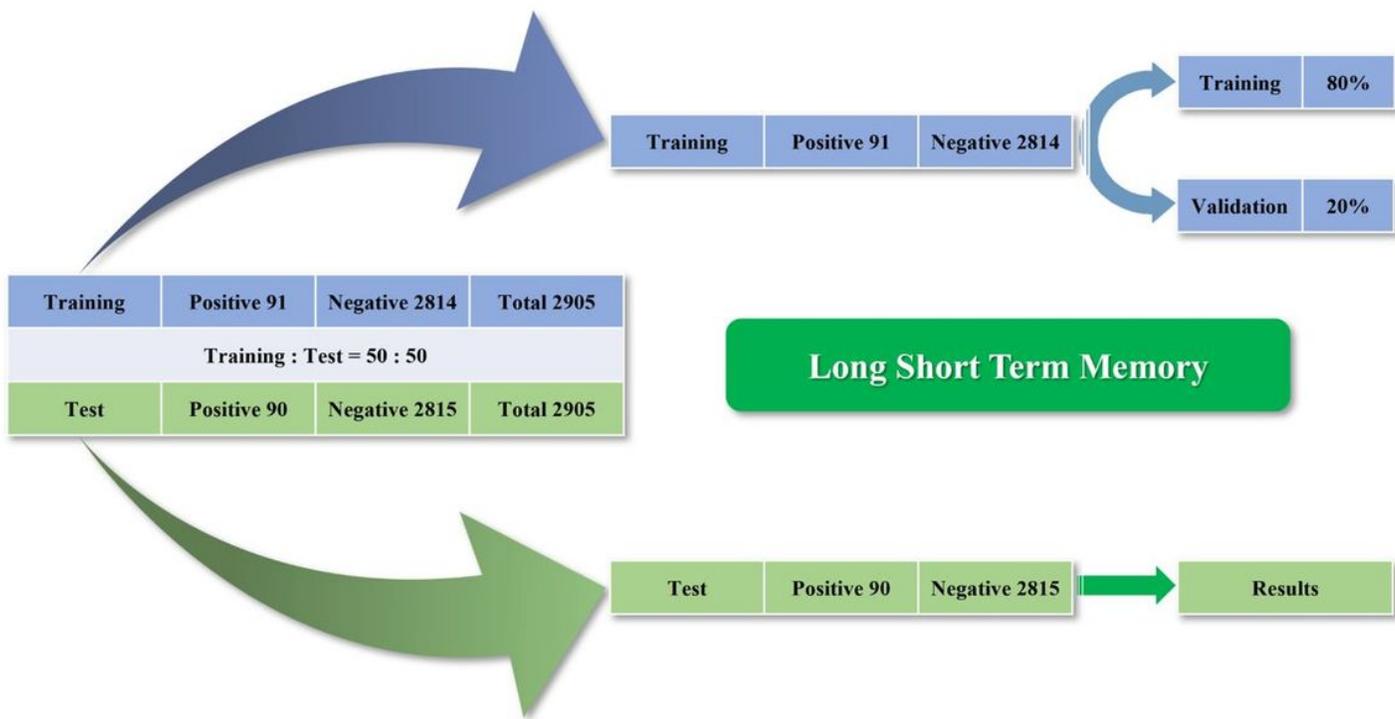27. Nemanja, S. M., Bratislav, B. P. & Milos, R. Multilayer long short-term memory (LSTM) neural networks in time series analysis.IEEE Xplore. Sep 2020https://doi.org/10.1109/ICEST49890.2020.9232710

Figures

Training	Positive 91	Negative 2814	Total 2905
Training : Test = 50 : 50			
Test	Positive 90	Negative 2815	Total 2905

Training	Positive 91	Negative 2814

Training	80%

Validation	20%

Long Short Term Memory

Test	Positive 90	Negative 2815

Results

Figure 1

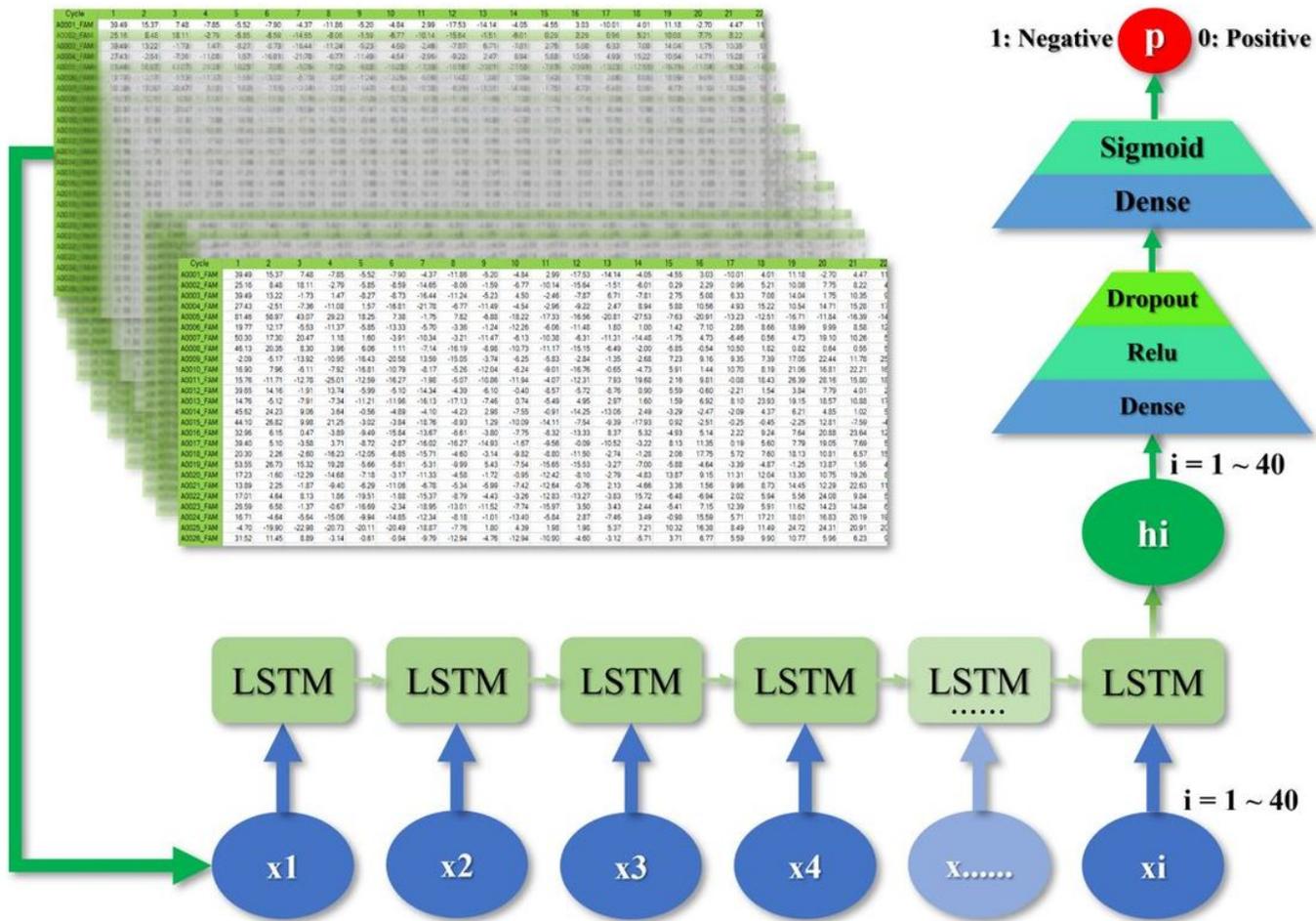Composition of the training and test dataset

Figure 2

Development of the deep learning model LSTM: Long short term memory
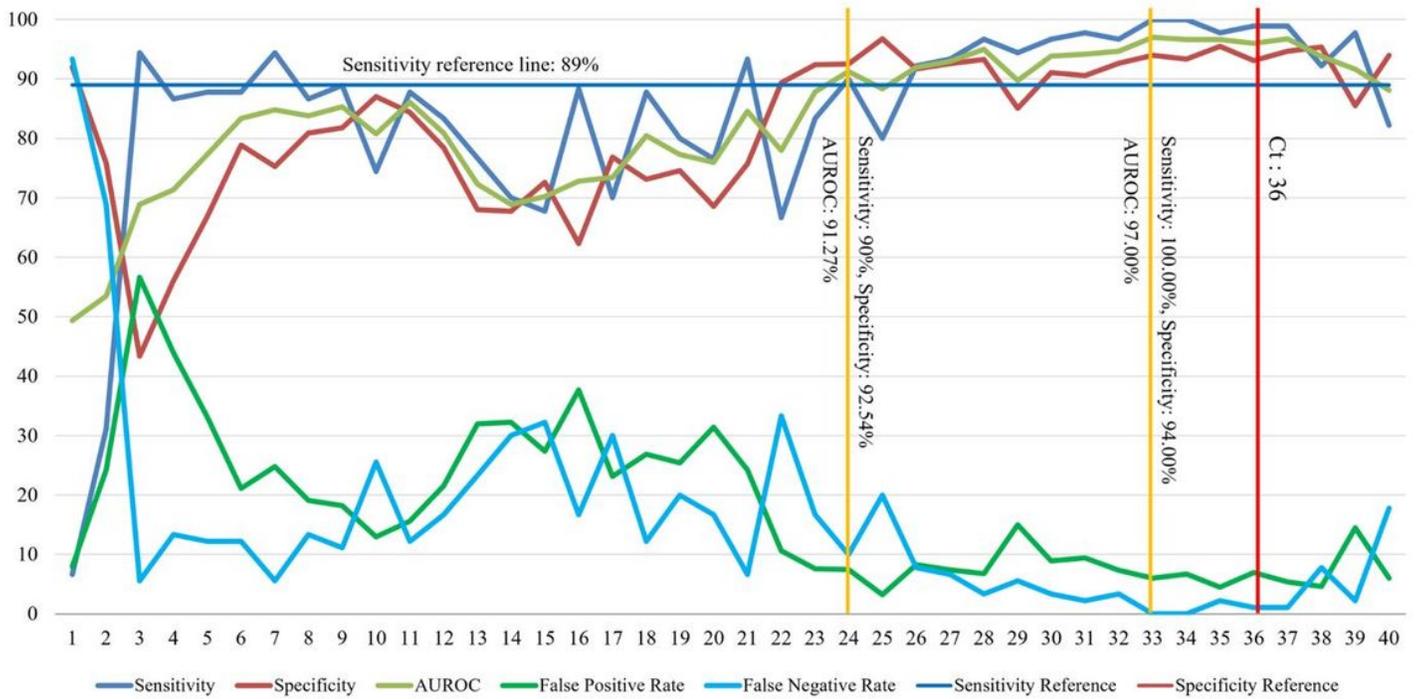
Figure 3

Comparison of the diagnostic performance of each model and RT-PCR AUROC: Area under the receiver operating characteristic curve; Ct: cycle threshold
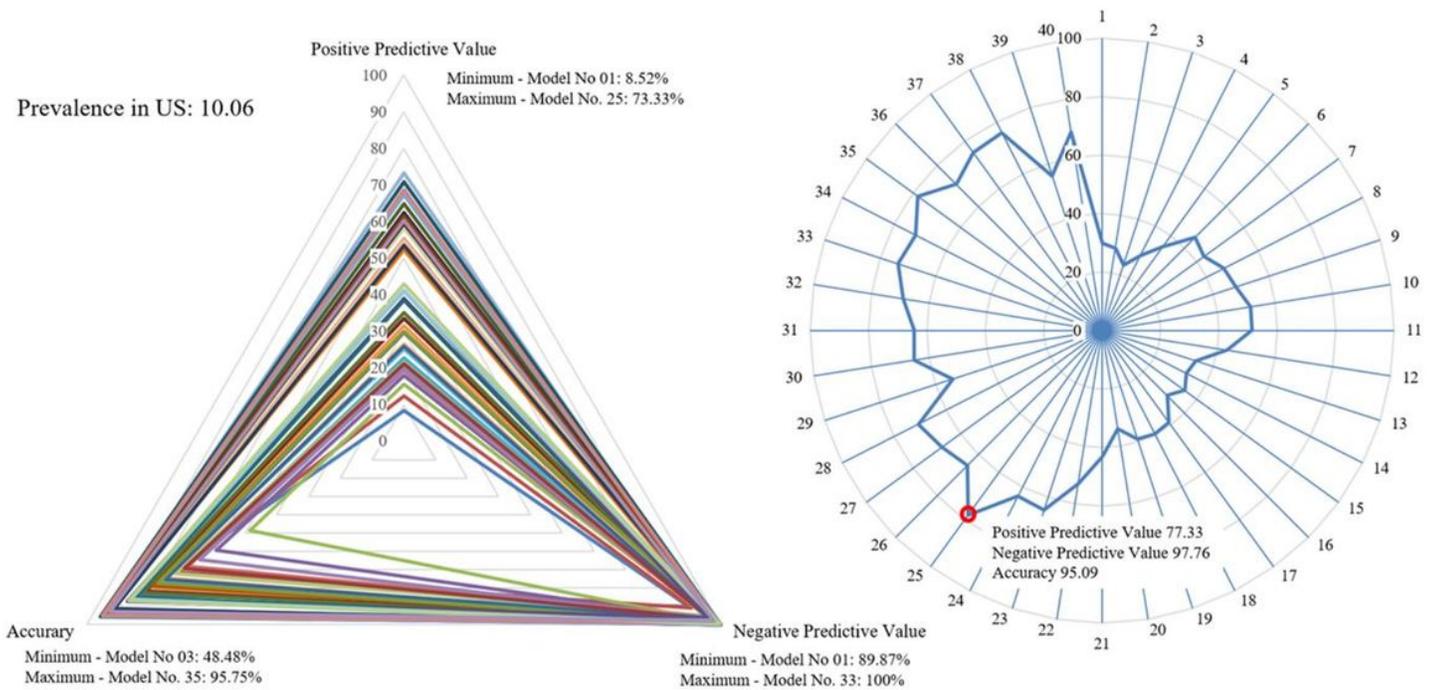

Figure 4

Comparison of each model considering screening performance affected by prevalence: United States Triangle radar chart: positive predictive values, negative predictive values and accuracy in each model; Polygonal radar chart: triangle radar chart area ratio in each model

Prevalence in Italy: 6.98

Positivie Predictive Value
Minimum - Model No 01: 5.92%
Maximum - Model No. 25: 65.02%

Accurary
Minimum - Model No 03: 46.94%
Maximum - Model No. 35: 95.68%

Negative Predictive Value
Minimum - Model No 01: 92.92%
Maximum - Model No. 33: 100%

Positive Predictive Value 65.02
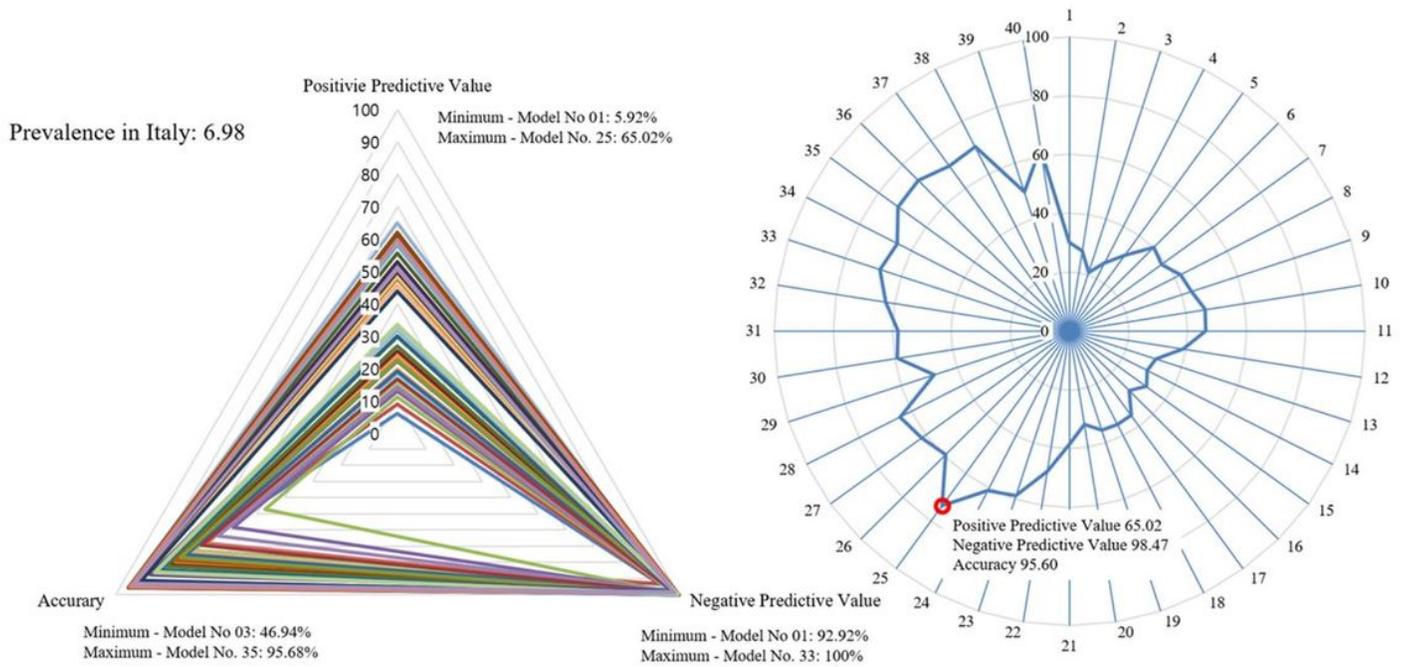Negative Predictive Value 98.47
Accuracy 95.60

Figure 5

Comparison of each model considering screening performance affected by prevalence: Italy Triangle radar chart: positive predictive values, negative predictive values and accuracy in each model; Polygonal radar chart: triangle radar chart area ratio in each model

Positive Predictive Value
Minimum - Model No 01: 0.23%
Maximum - Model No. 25: 6.43%

Prevalence in South Korea: 0.27

Accurary
Minimum - Model No 03: 43.52%
Maximum - Model No. 25: 96.72%

Negative Predictive Value
Minimum - Model No 01: 99.72%
Maximum - Model No. 33: 100%

Positive Predictive Value 65.02
Negative Predictive Value 98.47
Accuracy 95.60

Figure 6

Comparison of each model considering screening performance affected by prevalence: South Korea
Triangle radar chart: positive predictive values, negative predictive values and accuracy in each model;
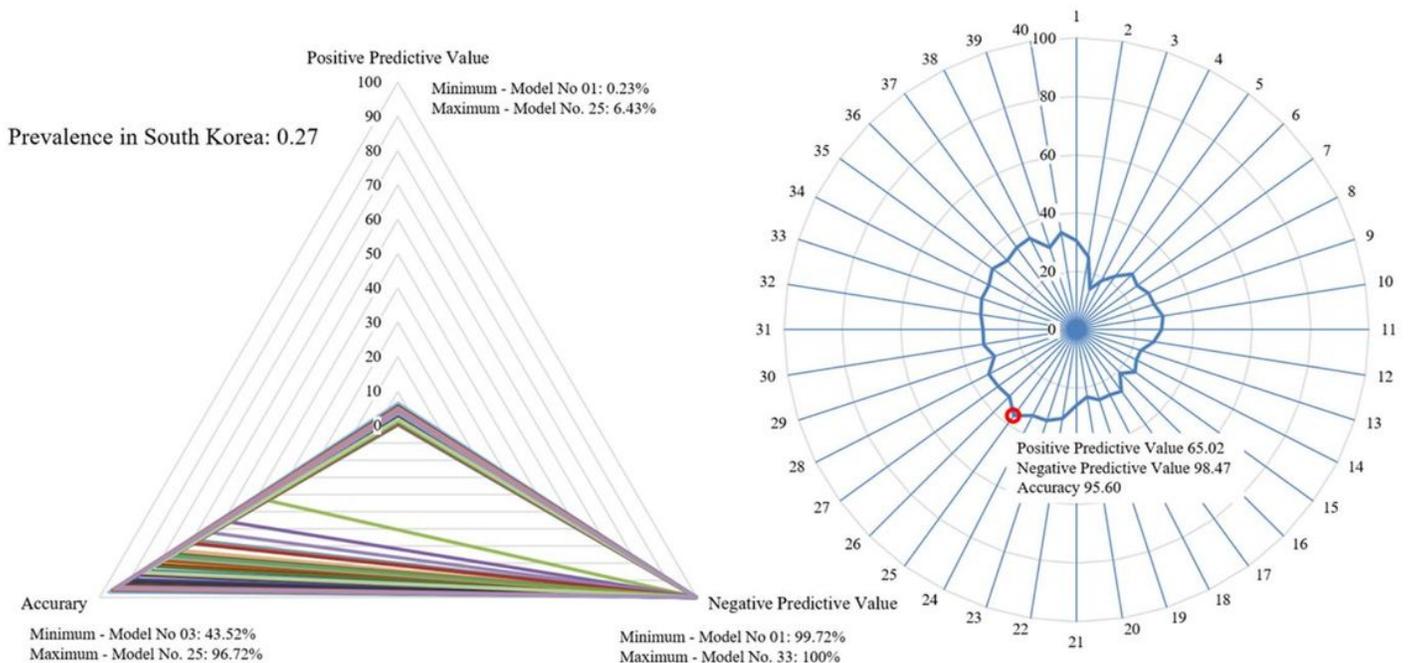Polygonal radar chart: triangle radar chart area ratio in each model

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.docx
- STARD2015Checklist.docx
- rawdata01.xlsx
- resultsdata.xlsx