An improved multivariate model that distinguishes COVID-19 from seasonal flu and other respiratory diseases

Xing Guo
Heping Hospital Affiliated to Changzhi Medical College
Yanrong Li
Changzhi Medical College
Xu Chang
Changzhi Medical College
Junfeng Li (lijunfengcz@163.com)
Heping Hospital Affiliated to Changzhi Medical College
Kefeng Li (kli@ucsd.edu)
University of California, San Diego

Short Report

Keywords: COVID-19; Multi-feature; Influenza; Random forest; Diagnostic model

Posted Date: May 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-28738/v1

Version of Record: A version of this preprint was published on October 21st, 2020. See the published version at https://doi.org/10.18632/aging.104132.

Abstract

Background: In many countries, the COVID-19 pandemic is occurring in the middle of flu season. Since the responses to COVID-19 are dramatically different, it is critical to accurately discriminate COVID-19 from seasonal flu and pneumonia caused by other common respiratory pathogens.

Methods: Fifty patients (eight patients with COVID-19, eight with influenza, and 34 with community-acquired pneumonia) were included in our study. Sixteen features, such as clinical symptoms, results of routine blood tests, first reverse transcription-polymerase chain reaction (RT-PCR), and chest CT, were collected. The importance of each feature in discriminating COVID-19 from others was ranked by the random forest algorithm. Models with single or multiple features were evaluated using receiver operating characteristic (ROC) curves, the F1 score, and Matthews correlation coefficient (MCC).

Results: An integrated multi-feature model (RT-PCR, CT features and blood lymphocyte percentage) yielded an area under the ROC curve of 0.97 (95% CI: 0.86 – 1, $P < 0.01$), an F1 score of 0.81 and an MCC of 0.78 in the training cohort as well as an F1 score of 0.86 and an MCC of 0.85 in the validation set.

Conclusion: The developed multivariate model showed better accuracy than the current nucleic acid-based method for the differentiation of COVID-19 from influenza and pneumonia caused by other common respiratory pathogens.

1. Introduction

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is a betacoronavirus. COVID-19 has spread globally since the first case reported in Wuhan, China, at the end of December 2019 [1]. As of April 16, 2020, COVID-19 has affected 185 countries and territories around the world, with 2,224,426 confirmed cases [2]. In addition to the ongoing COVID-19 pandemic, many countries are in the flu season. Patients with COVID-19 share many similarities in clinical symptoms and imaging characteristics with those with seasonal flu and pneumonia caused by other common respiratory pathogens [3]. Therefore, differentiating patients with COVID-19 from those with other seasonal respiratory diseases in a hospital setting is critical.

The accurate diagnosis of COVID-19 is challenging. Currently, reverse transcription-polymerase chain reaction (RT-PCR)-based analysis of nasopharyngeal swabs is the reference standard. However, the diagnostic sensitivity of RT-PCR testing is less optimal, and the magnitude of risk from false-negative test results is substantial [4]. Even though the utility of chest computed tomography (CT) for the detection of COVID-19 has been demonstrated in a few recent publications, many methodological flaws were present in these studies [5]. Additionally, several parameters in routine blood tests might be useful in predicting the severity of COVID-19 [6]. A new classification model that combines the advantages of various techniques is urgently needed.

Random forest (RF) has important advantages over other machine learning algorithms in terms of handling multidimensional and nonlinear biological data, the opportunity for efficient parallel processing, and its robustness to noise [7].

In this study, we developed an integrated multi-feature model based on RF to differentiate COVID-19 from seasonal flu and pneumonia caused by other common respiratory viruses. The performance of the model was validated using another cohort of subjects.

2. Materials And Methods

2.1 Subjects

This study was conducted on suspected COVID-19 patients who presented to Heping Hospital Affiliated to Changzhi Medical College from January 23, 2020, to February 20, 2020. All patients underwent CT scans on the day of admission. Pharyngeal swab samples were collected for RT-PCR analysis. Blood samples were collected from each participant. The age, sex, and clinical symptoms of the patients, as well as the epidemiological characteristics of COVID-19 were also collected. A total of 50 patients were enrolled in the study. The study protocol was approved by the Institutional Review Board of Heping Hospital Affiliated to Changzhi Medical College ((CMC-2020-1103), and written informed consent was obtained from each participant.

2.2 Blood tests

Routine blood tests, including white blood cell count (WBC), lymphocyte count (LYC), and blood lymphocyte percentage (LYP) (%), were performed on the blood samples.

2.3 Chest CT images

The CT examination was performed with a multislice spiral CT machine (TOSHIBA Aquilion 16, Japan) by two senior chest diagnostic radiologists who used a PACS workstation to read the axial images of standard 5 mm slice thickness, and 1 mm slice thickness images were used for multislice reconstruction to observe the lesions. The typical imaging manifestations of the patients were ground-glass opacification, consolidation, reticular shadow, and air bronchial sign, and most of the lesions were distributed in the subpleural areas. CT scans were read independently by two radiologists (blinded for review). Disagreements were resolved by a third experienced thoracic radiologist.

2.4 Data analysis

We followed the guidelines for the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). The analyses were performed in R 3.5.19. The importance of the features was ranked based on the mean decrease accuracy (MDA) using the random forest algorithm.

Once the ranked features were identified, several combinations of features were selected based on their ranking scores. We evaluated the performance of the models using the area under the receiver operating characteristic curve (AUROC). Classifier robustness was estimated by repeated double cross-validation (rdCV) and permutation testing 1,000 times. The accuracy, sensitivity, positive predictive value (PPV, %), negative predictive value (NPV, %), F1 score, and Matthews correlation coefficient (MCC) were also calculated. The optimal model was then validated using another set of patients.

3. Results

3.1 Patient characteristics

This study was conducted in a region outside of Wuhan, the epicenter of the COVID-19 outbreak. Out of 50 suspected patients enrolled, 8 had COVID-19, 8 had seasonal flu (influenza), and 34 had community-acquired pneumonia (CAP) (Table 1). There were no significant differences in characteristics among the three groups, including age, sex, clinical symptoms, WBC, and LYC. Compared to those with influenza and CAP, patients with COVID-19 had significantly lower LYP values ($P < 0.001$). Representative chest CT images for COVID-19, influenza and CAP are shown in Fig. 1.

3.2 An integrated model for the differentiation of COVID-19 from influenza and CAP

The capacity of each feature in discriminating COVID-19 from influenza and CAP was evaluated using a random forest model (decision trees = 1,000). The importance of the features was calculated based on MDA, and the top features are listed in Fig. 2.

The performance of the single and multi-feature models was assessed using ROC analysis and confusion matrices. Even though the first RT-PCR and CT alone had acceptable AUROCs (0.82 and 0.91), their F1 scores were only 0.66 and 0.67, respectively (Fig. 3A and 3B and Table 2); in addition, these two models had low Matthews correlation coefficient (MCC) values (0.67 and 0.64, respectively). We next evaluated the performance of the combination of multiple features. We found that an integrated model with the combination of LYP, RT-PCR, and CT performed well for distinguishing COVID-19 from seasonal flu and CAP, with an AUC of 0.97 (95% CI: 0.86 - 1, $P < 0.01$) (Fig. 3C). The permutation test showed that the integrated model was robust (Fig. 3D). Other classification metrics, including the accuracy, sensitivity, specificity, PPV, NPV, F1 score, and MCC, were also higher than those of the models with single features.

In the validation cohort, the multivariate model yielded an F1 score of 0.86, an MCC of 0.85, an accuracy of 96% (95% CI: 76.6 – 99.9), a sensitivity of 92% (95% CI: 89.4 -99.4) and a specificity of 88.2 (95% CI: 83.9 – 100).

4. Discussion

In this study, we successfully developed and validated a multivariate model that has the best performance for distinguishing COVID-19 from influenza and other respiratory diseases compared to

current approaches. Our model is particularly useful for screening COVID-19 in regions with a low incidence rate of COVID-19.

We used two additional classification metrics for the characterization of the developed models, including the F1 score and MCC. The F1 score is the weighted average of precision and recall. When the dataset is unbalanced (the number of patients in one group is much larger than the number of patients in the other groups), the traditional classifier accuracy is no longer a reliable metric [8]. In contrast, the F1 score takes the data distribution into account and is a better metric for our model in this study. Additionally, the Matthews correlation coefficient (MCC) is also a reliable metric in machine learning that produces a high score only if the prediction obtained good results in all four confusion matrix categories (true positives, false negatives, true negatives, and false positives) [8].

Nucleic acid testing is influenced by the specimen collector, sample source, and timing of the acquisition, resulting in a high false negative rate. Previous studies have reported that the positive rate of throat swab RT-PCR test is only approximately 59%, and the sensitivity is 30-60% [9]. In this study, three patients tested negative by RT-PCR four times before positive results were obtained the fifth time. In contrast, our integrated model using the combination of multiple features accurately identified these COVID-19 patients.

The study was limited to a small number of COVID-19 patients due to the low incidence rate in the non-epidemic area. The majority of the patients in our study had seasonal flu and CAP. Additional validation should be considered in a more diverse demographic group than our initial cohort prior to further clinical application.

In summary, we developed an integrated multivariate model that distinguishes COVID-19 from influenza and other respiratory diseases using the random forest algorithm. Compared to the current approaches, the new model may significantly reduce the possibility of false-negative and false-positive results for COVID-19.

Declarations

Acknowledgments

The authors would like to thank the nurses and patients involved in this study, without whom this work would have not been possible.

Funding

This project was funded by a special grant for COVID-19 prevention and control from Department of Education, Shanxi Province. The funder had no role in the design, execution, interpretation, or writing of the study.

Author information

Author contributions

X.G, J.L, and K.L. designed the study; X.G. Y. L., X. C. performed the experiments and collected the data; X.G, J.L, and K.L analyzed the data. X.G, Y.L., X.C., and K.Li wrote the original manuscript; All the authors revised the manuscripts; J.L. obtained the funding. X. G and Y. Li contributed equally to this work

Ethics declarations

Ethics approval and consent to participate

The study protocol was approved by the Institutional Review Board of Heping Hospital Affiliated to Changzhi Medical College (CMC-2020-1103), and the written informed consent was obtained from each participant.

Consent for publication

Not applicable.

Conflict of Interests

The authors declare that they have no competing interests.

Availability of data and materials

All data generated or analyzed during this study are included in this published article.

Competing interests

The authors declare no competing interests.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33; doi: 10.1056/NEJMoa2001017.
2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020; doi: 10.1016/S1473-3099(20)30120-1.
3. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA. 2020; doi: 10.1001/jama.2020.1585.
4. West CP, Montori VM, Sampathkumar P. COVID-19 Testing: The Threat of False-Negative Results. Mayo Clinic Proceedings. 2020; doi: 10.1016/j.mayocp.2020.04.004.
5. Hope MD, Raptis CA, Henry TS. Chest Computed Tomography for Detection of Coronavirus Disease 2019 (COVID-19): Don't Rush the Science. Ann Intern Med. 2020; doi: 10.7326/M20-1382.

6. Tan L, Wang Q, Zhang D, Ding J, Huang Q, Tang YQ, et al. Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study. Signal Transduct Target Ther. 2020;5(1):33; doi: 10.1038/s41392-020-0148-4.
7. Lebedev AV, Westman E, Van Westen GJ, Kramberger MG, Lundervold A, Aarsland D, et al. Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. Neuroimage Clin. 2014;6:115-25; doi: 10.1016/j.nicl.2014.08.023.
8. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6; doi: 10.1186/s12864-019-6413-7.
9. Yang Y, Yang M, Shen C, Wang F, Yuan J, Li J, et al. Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. medRxiv. 2020:2020.02.11.20021493; doi: 10.1101/2020.02.11.20021493.

Tables

Table 1 Patient characteristics

Characteristics	COVID-19 (n = 8)	Influenza (n = 8)	CAP (n = 34)	*P* value
Age (years)	25.1 (24.2 - 62.5)	29.5 (25.6 - 54.3)	31 (23.1 - 56.4)	0.93
Male (%)	4 (50%)	4 (50%)	22 (64.5%)	0.36
Fever (%)	2 (25%)	4 (50%)	15 (44.1%)	0.43
Cough (%)	2 (25%)	1 (12.5%)	12 (35.3%)	0.37
Sore throat (%)	2 (25%)	1 (12.5%)	3 (8.8%)	0.22
Fatigue (%)	2 (25%)	2 (25%)	4 (11.7%)	0.27
WBC (10^9/L)	5.3 (3.6 - 6)	4.9 (3.2 - 6.2)	5.5 (4.1 - 6.5)	0.65
Lymphocyte count (10^9/L)	0.93 (0.76 - 1.35)	1.4 (0.9 - 1.9)	1.5 (1.3 - 2.1)	0.08
Lymphocyte percentage (%)	14.2 ± 6.3	36.4 ± 7.1	33.9 ± 10.1	<0.001

The data are presented as the mean and standard deviation (SD) or median and IQR. The chi-square test or Kruskal-Wallis test was used. WBC: white blood cell count.

Table 2 The performance of the models trained by the random forest algorithm for COVID-19

Classification models	Model performance (Mean and 95% CI)						
	Accuracy (%)	F1 score	MCC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
PCR	92.0 (73.9 - 99.1)	0.66	0.67	100 (75.8 - 100)	91.3 (71.9 - 98.9)	50.1 (21.1 - 78.9)	100
CT	84.0 (63.9 - 95.5)	0.67	0.64	100 (79.8 - 100)	80.9 (58.1 - 94.6)	50.1 (29.3 - 70.7)	100
Integrated model-training set	92.0 (73.9 - 99.1)	0.81	0.78	100 (89.8 - 100)	90.5 (69.6 - 98.8)	86.7 (74.8 -92.2)	100
Integrated model-validation set	96 (79.6 - 99.9)	0.86	0.85	92 (89.4 - 99.4)	88.2 (83.9 - 100)	92 (85 – 100)	94.5 (79.4 - 99.1)

MCC: Matthews correlation coefficient; PPV (%): positive predictive value; NPV (%): negative predictive value.

Figures

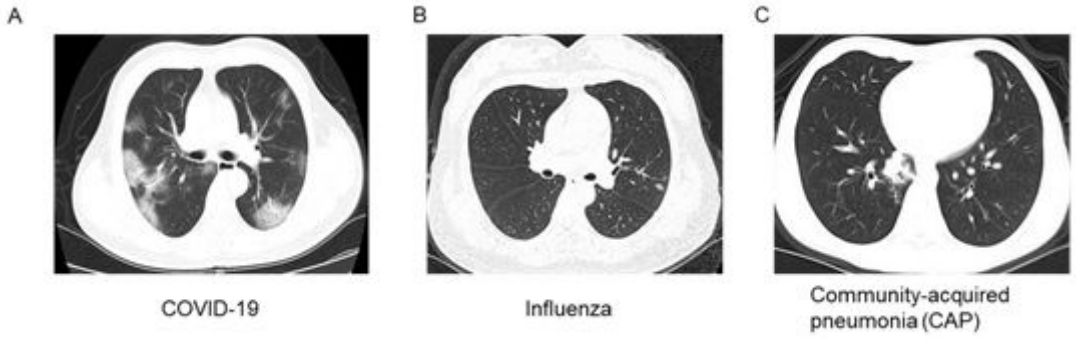

Figure 1

The representative chest images for patients with COVID-19 (A), influenza (B) and community-acquired pneumonia (CAP) (C).

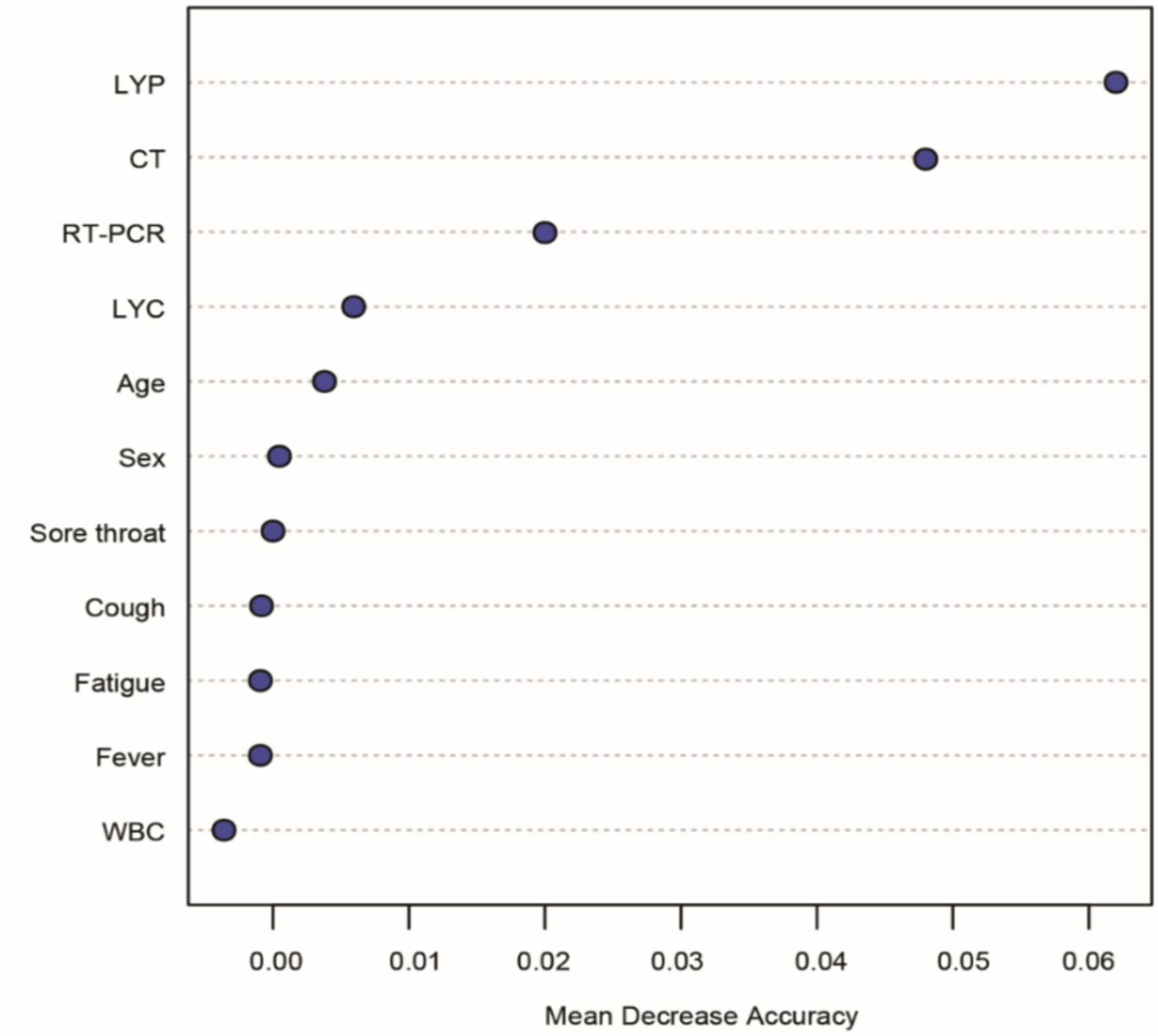

Figure 2

Features ranked by mean decrease accuracy (MDA) in the random forest model for classification between COVID-19 and other infections. Number of trees = 1000. LYP: lymphocyte percentage; WBC: white blood cell.

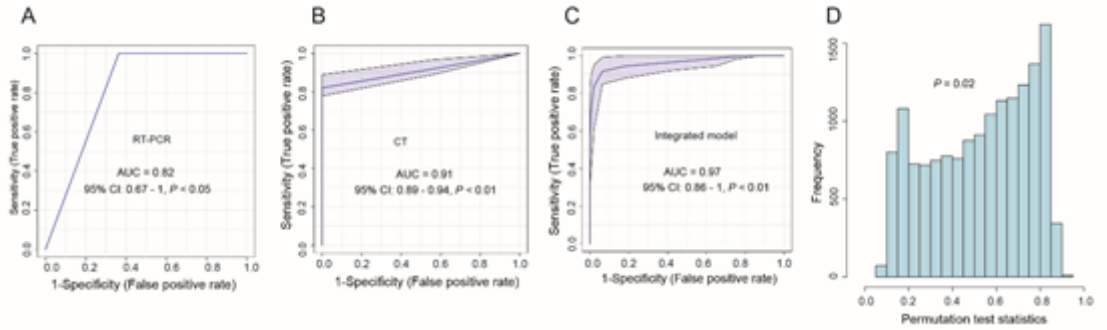

Figure 3

The development of an integrated model for the differentiation of COVID-19 from other respiratory diseases in the training set. (A) ROC curve for the performance of first RT-PCR; (B) ROC curve for the performance of CT; (C) ROC curve for the integrated model. The integrated model contained the results of first RT-PCR, CT, and LYP in the blood. (D) The cross-validation of the integrated model using the permutation test (1000 times).