OPEN ACCESS Freely available online

Journal of Antivirals & Antiretrovirals

Research Article

A Prediction Model Based on Machine Learning for Diagnosing the Early COVID-19 Patients

Nannan Sun[1], Ya Yang[2], Lingling Tang[3], Zhen Li[1], Yining Dai[4], Wan Xu[1], Xiaoliang Qian[1],Hainv Gao[3], Bin Ju[1*]

[1]Wowjoy Al lab, Hangzhou, China; [2]National Clinical Research Center for Infectious Diseases, The first Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China; [3]Department of Infectious Diseases, Shulan (Hangzhou) Hospital Affiliated to Zhejiang Shuren University Shulan International Medical College Hangzhou, China; [4]Department of Infectious Diseases, Zhejiang Provisional People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, China

ABSTRACT

Objective: To improve the timeliness for the early COVID-19 infection diagnosis, it is essential to develop a decision-making tool to assist early diagnosis of COVID-19 patients in fever clinics.

Materials and methods: This paper aims at extracting risk factors from clinical data of 912 early COVID-19 infected patients and utilizing four types of traditional machine learning approaches including Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and a deep learning-based method for diagnosis of early COVID-19.

Results: The results show that the LR predictive model presents a higher specificity rate of 0.95, an Area Under the receiver operating Curve (AUC) of 0.971 and an improved sensitivity rate of 0.82, which makes it optimal for the screening of early COVID-19 infection. We also perform the verification for generality of the best model (LR predictive model) among Zhejiang population, and analyse the contribution of the factors to the predictive models.

Discussions: Under the background of COVID-19 pandemic, the early diagnosis of COVID-19 still face severe challenges, a decision-making tool assisting early diagnosis of COVID-19 patients is vital for fever clinics.

Conclusions: Our manuscript describes and highlights the ability of machine learning methods for improving the accuracy and timeliness of early COVID-19 infection diagnosis. The higher AUC of our LR-base predictive model makes it a more conducive method for assisting COVID-19 diagnosis. The optimal model has been encapsulated as a mobile application (APP) and implemented in some hospitals in Zhejiang Province.

Keywords: Infectious diseases; Machine learning; Artificial intelligence; COVID-19; SARS-CoV-2

INTRODUCTION

The coronavirus disease 2019 (COVID-19) cases were first reported in Outbreak area in December 2019. Soon after, Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV-2), this new emerging virus has spread rapidly in over 200 countries and areas [1,2]. On March 11, 2020, the World Health Organization (WHO) declared that COVID-19 outbreaks a global pandemic. COVID-19 is a novel pathogen with characteristics of fast transmission and strong infectivity [3,4]. The early symptoms of COVID-19 are similar to other respiratory infectious diseases, which makes it difficult for early differential diagnosis [5-7]. So far, accurate RT-PCR test has been regarded as the gold standard for the diagnosis of COVID-19. However, RT-PCR tests are complicated in operation and it usually takes 5-6 hours or even longer to get the results [8]. Additionally, due to the low virus loads in early infected COVID-19 patients, RT-PCR tests show false negative results in a number of cases [9,10]. It has greatly hindered the prevention and control of the global pandemic. Thus, it is dramatically essential to establish a rapid diagnostic model to screen high-risk patients with COVID-19 infection.

In recent years, machine learning solutions are widely used to predict diagnosis and individual risk factors for diseases, and support clinical decisions [11]. Some machine learning methods have achieved remarkable results in medical filed, and a method with superior classification precision would provide better robustness for predicting unknown data [11-16]. Chhatwal et al. utilized LR to create a breast cancer risk estimation model based on the descriptors of National Mammography Database (NMD)

Correspondence to: Bin Ju, Hangzhou Wowjoy Information Technology Co., Ltd, Hangzhou, China, E-mail: bin.ju@wowjoy.cn

Received: April 01, 2021; **Accepted:** April 15, 2021; **Published:** April 22, 2021

Citation: Sun N, Yang Y, Tang L, Li Z, Dai Y, Xu W, et al. (2021) A Prediction Model Based on Machine Learning for Diagnosing the Early COVID-19 patients. J Antivir Antiretrovir. S18:002.

format that can aid in decision-making for early detection of breast cancer [17]. Pinto et al. used RF method to identify the patients who suffer from Alzheimer's disease based on ADNI datasets, which shows better accuracy and can be used as clinical assistant diagnosis [18]. Recently, the combination of machine learning approaches and epidemic infectious diseases has been emerged extensively. Hong et al. used SVM with double class analysis for MERS-COV epidemiological study and discovered the relevance between two sequences of MERS-COV [19]. Jia et al. constructed predictive model with higher accuracy for antigen mutation of influenza virus subtype H1, which used CART DT algorithm combined with amino acid variation sites of viral proteins [20]. The combination of machine learning and medical data has become the main development direction to meet the needs of early diagnosis and prognosis assessment.

In this study, we attempted to identify the best appropriative algorithm for early COVID-19 detection based on clinical big data. We analysed clinical data of 912 patients who were confirmed as COVID-19 or other respiratory infectious diseases from 18 hospitals in Zhejiang Province, focusing on extraction of risk factors and construction of five types of classification models: SVM, LR, DT, RF as well as Deep Neural Network (DNN). Four epidemiological factors and six clinical manifestations were selected by feature engineering approach as diagnostic models input, and they were much fewer than candidate features of medical records. Essentially, the diagnostic model constructed with fewer meaningful clinical factors is practical for outpatient service. Clinical symptoms, laboratory tests and imaging findings play significant roles in identification of COVID-19 infection [21]. To evaluate the contributions of clinical symptoms, laboratory tests and imaging information for diagnostic models, we established predictive models based on the data excluding epidemiological information. It was found that the diagnostic models established with clinical symptoms, laboratory tests and imaging information only presented poorer performance. In other words, epidemiological information tremendously affects the performance of COVID-19 predictive models. Briefly, making full use of clinical manifestations and epidemiological characteristics integrated is essential for constructing the early diagnosis model of COVID-19.

MATERIALS AND METHODS

Data construction

The COVID-19 dataset contains clinical information of 914 suspected patients who were from 18 hospitals in Zhejiang between Jan 17 and Feb 19, 2020. Suspected cases were diagnosed according to the 5th edition of Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment. COVID-19 should be suspected if subjects conform to any one of the criteria in the epidemiological history and any two of the standards in clinical presentations. If there is no epidemiological history, suspected cases should meet three of the criteria in clinical presentations [22].

Considering about the completeness of the clinical information, we firstly screened out the patients with complete clinical records, which results in total number of 912 eligible patients. We then split processed patients into training (80%) and validation (20%) partitions randomly to train our models. Subsequently, we collected 115 clinical dataset from other hospitals in Zhejiang as test partition to verify the universality of implemented models in Zhejiang population.

To obtain the datasets for early stage COVID-19 rapid diagnostic models, all selected suspected patients were categorized into positive or negative cases. The patients who met any one of the following criteria's were considered to be positive cases.

Positive RT-PCR test results in throat swab, sputum, blood samples. The genetic sequences detected in the samples are highly homologous to the known SARS-CoV-2.

Positive cases are considered to be the patients confirmed as COVID-19 infection by RT-PCR. Conversely, negative cases are patients excluded as COVID-19 infection by RT-PCR for at least two times. The 912 eligible participants enrolled in this study, include 361 COVID-19 infected patients (positive cases) and 551 COVID-19 non-infected patients (negative cases). Each patient's clinical record contains 31 factors including gender, age, coexisting diseases, epidemiological information, laboratory tests, clinical symptoms and imaging findings. Details of these 31 factors and their distribution characteristics on training and validation dataset are shown in Table 1.

Table 1: The characteristics of positive and negative samples on the training set and validation set.

Factors	Training data set		Validation data set	
	Positive(n=293)	Negative(n=436)	Positive(n=68)	Negative(n=115)
Gender (Female)	127	213	31	62
Age(year)	47.39 ± 14.38	38.53 ± 18.14	46.18 ± 14.73	35.33 ± 16.85
Coexisting diseases	102	59	19	17
Travel or residence history over the past 14 days				
Outbreak area	97	58	23	12
Neighboring areas of outbreak area	7	53	2	15
Other areas with persistent local transmission or community with definite cases	136	212	26	54

Exposure to patients with fever or respiratory symptoms over the past 14 days who had a travel or residence history				
Outbreak area	77	50	23	14
Neighboring areas of outbreak area	7	18	1	6
Other areas with persistent local transmission or community with definite cases	79	22	20	10
Suspected patients	4	6	0	5
Relationship with a cluster outbreak	104	13	29	3
Exposure to wildlife	0	1	1	0
Contact with patients of influenza A	1	12	2	3
Contact with patients of influenza B	2	11	3	4
Body temperature	37.54 ± 0.852	37.48 ± 0.848	37.39 ± 0.69	37.46 ± 0.73
Dry cough	156	128	32	55
Sputum	107	104	23	31
Factors	Training data set		Validation data set	
	Positive(n=293)	Negative(n=436)	Positive(n=68)	Negative(n=115)
Fatigue	81	46	17	13
Dyspnea	29	10	4	1
Conjunctival congestion	1	2	1	2
Nasal congestion	13	37	1	9
Diarrhea or bellyache	31	17	6	4
Dizziness or headache	24	38	4	9
Nausea or vomiting	10	5	3	2
Sore throat	15	47	4	14
Muscle soreness	16	3	1	0
White blood cell count (× 10^9)	5.47 ± 2.63	7.36 ± 3.04	5.10 ± 1.88	7.08 ± 2.89
Lymphocyte count (× 10^9)	1.22 ± 0.88	1.68 ± 0.86	1.16 ± 0.54	1.81 ± 0.78
Neutrophil cell count (× 10^9)	4.01 ± 4.67	5.12 ± 3.47	3.47 ± 1.68	4.58 ± 2.40
C-reactive protein level (mg/L)	21.77 ± 27.79	18.72 ± 35.26	17.33 ± 20.97	17.15 ± 34.93
Imaging changes of Chest X-ray or CT				
Normal	10	194	6	48
Unilateral local patchy shadowing	94	104	13	28
Bilateral multiple ground glass opacity	94	77	24	20
Bilateral with pulmonary consolidation	84	24	25	6
Other imaging alterations	11	37	0	13

Feature selection

Feature selection is used to select effective factors from numerous features to reduce the feature space dimension and classification error rate. We leveraged embedded feature engineering approach based on logistic regression algorithm with L2 penalty to select COVID-19 risk factors from the 31 factors mentioned above. Finally, 10 factors were chosen for the early COVID-19 prediction task by setting the threshold as 0.85. The final selected factors include four epidemiological features (relationship with a cluster outbreak, travel or residence history over the past 14 days in Outbreak area, exposure to patients with fever or respiratory symptoms over the past 14 days who had a travel or residence history in outbreak area, exposure to patients with fever or respiratory symptoms over the past 14 days who had a travel or residence history in other areas with persistent local transmission, or community with definite cases) and six clinical manifestations(muscle soreness, dyspnea, fatigue, lymphocyte count (× 10^9/L), white blood cell count (× 10^9/L) and imaging changes of Chest X-ray or CT). In practice, the diagnostic model constructed with fewer incoherent factors is beneficial and practical for outpatient service. Details of selected risk factors and their related coefficients are shown in Table 2. The importance of the factors relies on absolute value of the coefficients. Table 2 suggests that imaging changes of Chest X-ray or CT is more vital than others. Table 2 also shows that the tolerance of these 10 factors is more than 0.1 and variance inflation factor of them is less than 10, which indicated that there was no collinearity among selected factors.

Table 2: The coefficients, tolerance and variance inflation factor of factors selected by feature engineering.

Factors	Coefficients	Tolerance	VIF factor
Imaging changes of Chest X-ray or CT	3.15	0.42	2.38
Relationship with a cluster outbreak	2.44	0.42	2.38
Travel or residence history over the past 14 days in outbreak area	1.58	0.75	1.33
Muscle soreness	1.46	0.94	1.06
Exposure to patients with fever or respiratory symptoms over the past 14 days who had a travel or residence history in other areas with persistent local transmission, or community with definite cases	1.07	0.75	1.34
Dyspnea	1.07	0.9	1.11
Fatigue	0.97	0.79	1.26
Factors	Coefficients	Tolerance	VIF Factor
Exposure to patients with fever or respiratory symptoms over the past 14 days who had a travel or residence history in outbreak area	0.87	0.75	1.34
Lymphocyte count(× 10^9/L)	-1.36	0.25	4.02
White blood cell count(× 10^9/L)	-3.05	0.23	4.41

Methodology

In this study, we conduct four conventional types of machine learning algorithms and a deep learning solution to establish the early stage COVID-19 rapid diagnostic models. We implement LR model with L2 regularization penalty, and train other three models including SVM with kernel of rbf, ID3 DT and RF. The RF model is constructed by 50 decision trees with information gain algorithm. This study used deep learning-base method, namely DNN, which is a four-layer network with the hidden dimension of 64, 32, 16 and 20 respectively. A Softmax layer is added at the top of the network to output the probability of a patient infected with COVID-19.

We evaluate the performance of the early stage COVID-19 diagnostic models at the 20% validation using familiar assessment strategies, which include measuring accuracy and the AUC generated by plotting sensitivity vs. 1-specificity. Classification accuracy is obtained via an optimum cut-off point. AUC measures the overall performance of the recall concerning different false positive rate, which exhibits robustness for performance assessment of predictive models [23]. Models with higher AUC will show more powerful identified and diagnostic capacities to assist health care workers. High-sensitivity (or recall rate of positive cases) and high-specificity (or recall rate of negative cases) play a vital role in screening the infectious patients [24]. Essentially, a model with high sensitivity can correctly identify patients infected with COVID-19 for timely treatment, while a model with high specificity can excellently screen non-infective patients, thereby effectively avoid cross infection.

RESULTS

The experiments we conduct to evaluate the performance of the five types of predictive models are illustrated in this section. We evaluate the predictive models on validation set and compare the results of validation to obtain the best solution for identifying early COVID-19 infection. Ultimately, we test the best model based on test dataset to obtain general diagnostic model for Zhejiang population.

We implement multiple model structures as our constructed models and deploy different combinations of feature inputs (Table 3). Table 4 summarizes the performance of conventional solutions and deep learning-based methods. Table 3 part (a) reveals the performances of predictive models constructed based on the raw dataset including 31 factors (Table 1), and part (b) exhibits the performances of models established with ten factors selected by using feature selection approach. Feature selection is intended for data dimensionality reduction [4]. In practice, the diagnostic models constructed with less meaningful clinical factors are more practical for outpatient services. The results in Table 3 demonstrate that the predictive models of part (b) perform slightly better than that of part (a) in terms of AUC. The sensitivity, specificity, as well as accuracy of these predictive models of part (b) are relatively approximate to those of part (a). Thus, feature selection partly improves the performances of COVID-19 diagnostic models, and the ROC curve of some selected high performing machine learning models are shown in Figure 1. Table 4 part (b) shows that LR combined with feature selection outperforms other four methods by reaching an AUC of 0.971, high-specificity of 0.95 and accuracy of 0.90 respectively. These results suggest that the combination of LR and feature selection approach presents the best AUC and specificity among five categories of classification methods. Higher specificity of model will facilitate the elimination of infected diseases such as COVID-19 infection. In addition, according to the clinical experience of experts, the AUC (0.86) calculated by the diagnostic scale is compared, and the LR diagnostic model shows better performance. Therefore, LR can be selected as the optimum classification model for the early-stage COVID-19 rapid screening.

Table 3: The performance comparison of various machine learning models on validation set with different sets of features.

Model	Sensitivity	Specificity	Accuracy	AUC
(a) Models constructed on raw dataset				
LR	0.84	0.95	0.91	0.967
SVM	0.74	0.87	0.82	0.916
DT	0.81	0.89	0.86	0.851
RF	0.82	0.9	0.87	0.957
DNN	0.84	0.88	0.86	0.946
(b) Models constructed combination with feature selection				
LR+FS	0.82	0.95	0.9	0.971
SVM+FS	0.82	0.94	0.9	0.952
DT+FS	0.85	0.9	0.88	0.884
RF+FS	0.85	0.88	0.87	0.957
DNN+FS	0.87	0.89	0.86	0.953
(c) Models constructed on the dataset excluding epidemiological information				
LR-exclude	0.23	0.23	0.23	0.23
epidemiology	0.76	0.85	0.82	0.872
SVM-exclude epidemiology	0.71	0.87	0.81	0.871
DT-exclude epidemiology	0.71	0.83	0.79	0.761
RF-exclude epidemiology	0.69	0.87	0.8	0.875
DNN-exclude epidemiology	0.71	0.89	0.82	0.854
(d) Models combination with feature selection on the dataset excluding epidemiological information				
LR+FS-exclude epidemiology	0.76	0.87	0.83	0.871
SVM+FS-exclude epidemiology	0.81	0.87	0.85	0.889
DT+FS-exclude epidemiology	0.71	0.88	0.81	0.815
RF+FS-exclude epidemiology	0.71	0.86	0.8	0.848
DNN+FS-exclude epidemiology	0.76	0.84	0.81	0.864

Note: FS: Feature Selections exclude epidemiology: models established on the data excluding epidemiology information.

Table 4: The sensitivity, specificity, accuracy and AUC for logistic regression on test dataset.

Model	Sensitivity	Specificity	Accuracy	AUC
LR+FS	0.87	0.95	0.91	0.95

Note: LR: Logistic Regression; FS: Feature Selection

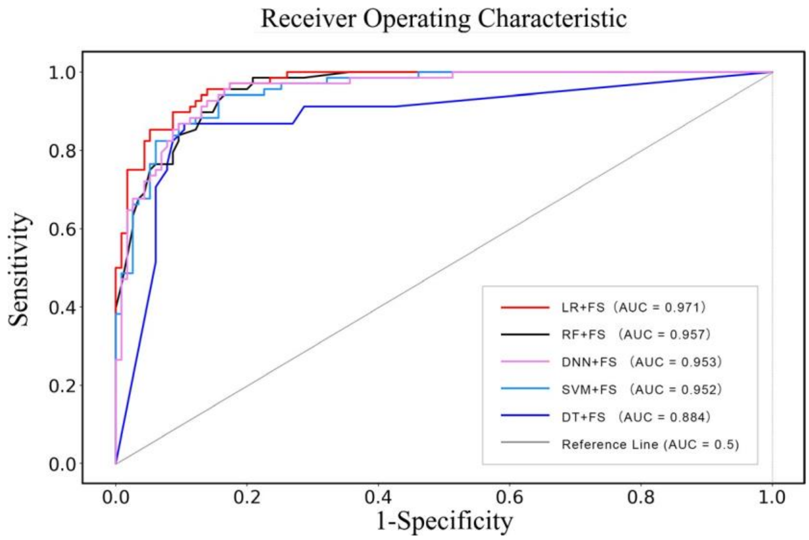

Figure 1: ROC curve of chosen high performing machine learning models. FS: Feature Selection.

Under the background of COVID-19 pandemic, clinical symptoms, laboratory tests and imaging findings are vital clinical criterion for the diagnosis of COVID-19 infection. In order to verify the contribution of above-mentioned three indicators to the COVID-19 diagnostic models, we establish predictive models based on the dataset excluding epidemiological information. The performances of various predictive models are shown in Table 3 (c) and (d). Results in part (a) and part (b) illustrates that epidemiological information is beneficial for early COVID-19 rapid diagnostic models construction. In the absence of epidemiological information, the sensitivity, specificity and accuracy of the predictive models (part (c)) exhibits sharp reduction compared with the models shown in part (a). In addition, the five types of machine learning approaches combining with feature selection are constructed based on the dataset excluding epidemiological information, as is shown in Table 4 part (d). Compared with part (c), AUC of part (d) is slightly improved. While due to the absence of epidemiological information, part (c) and part (d) show poorer performances compared with part (a) and part (b). In brief, it indicates laterally that epidemiological information is essential for constructing the early COVID-19 diagnostic models in Zhejiang population.

The above results clearly illustrate that the combination of traditional logistic regression method and feature selection has a great probability to predict early COVID-19 infection. And construction of highly precious diagnostic model relies on integrating and taking the most advantages of clinical symptoms, laboratory tests, imaging findings as well as epidemiological information.

Moreover, LR algorithm is proved as the most ideal method among the five classification solutions for the early COVID-19 rapid screening. The experiment performed in this study uses test dataset for verifying generality of the optimum diagnosis model. As is shown in Table 4, the sensitivity, specificity, accuracy and AUC of the LR+FS model on test dataset are 0.87, 0.95, 0.91 and 0.95, respectively. These results show that the predictive model constructed by combination of logistic regression and feature selection as early COVID-19 rapid diagnostic tool is universally applicable in Zhejiang Province.

DISCUSSION

Under the background of COVID-19 pandemic, the early prevention and control of COVID-19 still face severe challenges. According to the reports, the most common early symptoms of COVID-19 are fever, cough, fatigue, and myalgia, followed by diarrhoea, nausea, headache and sore throat [25,26]. As the disease goes on, some infected patients, especially those with low immune functions, gradually become dyspnea [21,27]. Additionally, complications such as acute arrhythmia and shock, Acute Respiratory Distress Syndrome (ARDS), are probably related to a poor prognosis [28,29]. Thus, early prediction of suspected patients and early aggressive treatment of confirmed patients is the key to reduce cross infection and mortality. CT scan has become the main auxiliary tool for screening of COVID-19 cases. However, CT scan cannot be used to identify specific viral infections [30]. Moreover, some COVID-19 patients can also present with normal pulmonary imaging in early stage [31]. Clinical symptoms and laboratory tests are sometimes non-specific for early COVID-19 infection [21,32]. At present, RT-PCR is still the accepted detection method for the diagnosis of COVID-19 infection. While the time consuming and instability of test results are still the most struggling problems [8]. Therefore, to improve the timeliness for the early COVID-19 infection diagnosis, it is essential to develop a decision-making tool to assist early diagnosis of COVID-19 patients in fever clinics.

Current studies which analyse symptoms and laboratory examination results of COVID-19 patients mainly focus on predicting mortality risk and progression of the disease [33]. Only few studies aim at COVID-19 early diagnosis. At present, Meng et al. selected nine representative variables (including age, activated partial thromboplastin time, red blood cell distribution width-sd, uric acid, triglyceride, serum potassium, albumin/globulin, 3-hydroxybutyrate, serum calcium) and constructed an optimized diagnostic model through Lasso regression screening and Multivariate logistic regression based on 431 samples [34]. The AUC of their early COVID-19 screening model in the testing set and independent validation cohort were 0.890 and 0.872. Feng et al. used logistic regression with Lasso regression for features selection and screening model development based on clinical data of 132 recruited patients [35]. The final chosen features include 1 demographic variable (age); 4 variables of vital signs (e.g., Temperature (TEM), Heart rate (HR), etc.); 5 variables of blood routine values (e.g., Platelet count (PLT), Monocyte ratio (MONO%), Eosinophil count (EO#), etc.); 7 variables of clinical signs and symptoms (e.g., Fever, Fever classification, Shiver, etc.); and 1 infection-related biomarker (Interleukin-6 (IL-6)). The performance of their model constructed based on the final selected features in held-out testing set and validation cohort resulted in AUCs of 0.841 and 0.938, and specificity of 0.727 and 0.778. In our study, we selected four epidemiological features and six clinical manifestations from the raw dataset including 31 factors, further developed multiple models with various machine learning algorithms and screened an optimum early COVID-19 diagnostic model with an AUC of 0.971. We tested the best model based on LR on the external test data set, and its AUC and specificity were 0.950 and 0.95, respectively. Compared to previous studies, we screened out fewer risk factors based on a larger clinical data set, and the early COVID-19 diagnostic model we established has better performance and is more suitable for clinical assisted diagnosis. Moreover, our study is based on a large clinical data set, including a total of 912 patients who were confirmed to have early COVID-19 infection or other respiratory infectious diseases, which may contribute to mining more potential clinical information and improve generalization ability of diagnostic models. Considering the indisputable role epidemiological features play in the diagnosis of infectious diseases in clinic Zhang et al. we specifically studied the role of epidemiological information in diagnostic models. We found that the lack of epidemiological information greatly affected the accuracy, specificity and sensitivity of the model. It means that epidemiological information is vital for building an accurate COVID-19 diagnostic tool, and makes the utility and reliability of the previously reported diagnostic models questioned [36].

Nevertheless, this study still has several limitations. First of all, the recruited participants are limited to Zhejiang Province, which causes certain regional restrictions in the application of the predictive models. Further extremely concerning about the epidemiological characteristics and nationwide studies are needed to access the generality of the suggested model. Secondly, there is a lack of information on the progression and prognosis of COVID-19 as well as asymptomatic infection cases. Finally, more information of infections should be recruited to improve the accurate of screening model.

CONCLUSION

In our study, ten representative factors with significant identification value were selected and constructed diagnostic models. The model established an algorithm based on logistic regression can be used as a simple, fast, and effective tool for diagnosing the early COVID-19 infection with significant clinical value.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the grant from the emergency project of key research and development plan in National Science and Technology Major Project (2017ZX10204401) and Zhejiang Province (2020C03123-1; 2020C03123-2).

Competing interests

The authors declare that they have no competing interests.

Author contributions

Bin Ju conceived and designed the experiments, analysed the data, authored or reviewed drafts of the paper, and approved the final draft.

Nannan Sun performed the experiments, analysed the data, prepared figures and or tables, and approved the final draft.

Ya Yang authored or reviewed drafts of the paper, and approved the final draft.

Lingling Tang revised drafts of the manuscript, and approved the final draft.

Zhen Li revised drafts of the manuscript, and approved the final draft.

Hainv Gao revised drafts of the manuscript, and approved the final draft.

Yining Dai provided samples dates, and approved the final draft.

Nannan Sun and Ya Yang contributed equally to this work.

Human ethics

This study was approved by the Ethics Committee of Zhejiang Provincial People's Hospital (2020KY006). Exempt informed consent was approved because the subjects would not be exposed to any risk in this observational study, and the information of subjects was anonymized at collection and prior to analysis.

REFERENCES

1. Wu J, Liu J, Li S, Peng Z, Xiao Z, et al. Detection and analysis of nucleic acid in various biological samples of COVID-19 patients. Travel Med and Infect Dis. 2020;37:101673.

2. Xu XW, Wu XX, Jiang XG, Xu KJ, Ying LJ, Ma CL, et al. Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: Retrospective case series. BMJ. 2020;368.

3. Chan JF, Yuan S, Kok KH, To KK, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. Lancet. 2020;395(10223):514-523.

4. Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. J Adv Res. 2020;24:91-98.

5. Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: What lessons have we learned? Int J Epidemiol. 2020;49(3):717-726.

6. Hon KL, Leung KKY, Leung AKC, Sridhar S, Qian S, Lee SL, et al. Overview: The history and pediatric perspectives of severe acute respiratory syndromes: Novel or just like SARS. Pediatr Pulmonol. 2020;55(7):1584-1591.

7. Chavez S, Long B, Koyfman A, Liang SY. Coronavirus disease (COVID-19): A primer for emergency physicians. Am J Emerg Med. 2020.

8. Samson R, Navale GR, Dharne MS. Biosensors: Frontiers in rapid detection of COVID-19. 3 Biotech. 2020;10(9):1-9.

9. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for typical 2019-ncov pneumonia: Relationship to negative RT-PCR testing. Radiology. 2020;296(2):E41-E45.

10. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: Summary of a report of 72 314 cases from the Chinese center for disease control and prevention. JAMA. 2020;323(13):1239-1242.

11. Yadav M, Perumal M, Srinivas M. Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos Solitons Fractals. 2020;139:110050.

12. Chao CM, Yu YW, Cheng BW, Kuo YL. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. J Med Syst. 2014;38(10):106

13. Yeom S, Giacomelli I, Menaged A, Fredrikson M, Jha S. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. J Comp Securi. 2019;1-36

14. Oliveira BL, Godinho D, O'Halloran M, Glavin M, Jones E, Conceição RC. Diagnosing breast cancer with microwave technology: Remaining challenges and potential solutions with machine learning. Diagnostics. 2018;8(2):36.

15. Luo ST, Cheng BW. Diagnosing breast masses in digital mammography using feature selection and ensemble methods. J Med Syst. 2012;36(2):569-577.

16. Fan CY, Chang PC, Lin JJ, Hsieh JC. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl Soft Comput. 2011;11(1):632-644.

17. Chhatwal J, Alagoz O, Lindstrom MJ, Kahn Jr CE, Shaffer KA, Burnside ES. A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. Am J Roentgenol. 2009;192(4):1117-1127.

18. Maggipinto T, Bellotti R, Amoroso N, Diacono D, Donvito G, Lella E, et al. DTI measurements for Alzheimer's classification. Phys Med Biol. 2017;62(6):2361.

19. Hong S, Choi S, Kim D, Yoon T. Epidemiological analysis of MERS-CoV using NN and SVM in respect to applicability of AI in multiple classes. IEEE. 2017:894-899.

20. Wang J, Wu YL. Prediction of antigenic variation of influenza virus subtype H1 based on machine learning. Inform Commun. 2018:63-64

21. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. Lancet. 2020;395:507-513.

22. Wang F, Liu J, Zhang P, Jiang W, Zhang L, Zhang M, et al. Expert

consensus on prevention and control of COVID-19 in the neurological intensive care unit (first edition). Stroke Vasc Neurol. 2020.5(3):242-249.

23. Belouch M, El Hadaj S, Idhammad M. Performance evaluation of intrusion detection based on machine learning using Apache Spark. Proced Comput Sci. 2018;127:1-6.

24. Lin YW, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PloS One. 2019;14(7):e0218942.

25. Han C, Duan C, Zhang S, Spiegel B, Shi H, Wang W, et al. Digestive symptoms in COVID-19 patients with mild disease severity: Clinical presentation, stool viral RNA testing, and outcomes. Am J Gastroenterol. 2020. 115(6):916-923.

26. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. Lancet Respiratory Med. 2020;8(4):420-422.

27. Koff WC, Williams MA. COVID-19 and immunity in aging populations-a new research agenda. N Engl J Med. 2020;383(9):804-805.

28. Hu Y, Sun J, Dai Z, Deng H, Li X, Huang Q, et al. Prevalence and severity of corona virus disease 2019 (COVID-19): A systematic review and meta-analysis. J Clin Virol. 2020:104371.

29. Li Y, Xia L. Coronavirus disease 2019 (COVID-19): Role of chest CT in diagnosis and management. Am J Roentgenol. 2020;214(6):1280-1286.

30. Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiol. 2020;295(1):202-227.

31. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497-506.

32. Wynants L, Van-Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ. 2020;369.

33. Meng Z, Wang M, Song H, Guo S, Zhou Y, Li W, et al. Development and utilization of an intelligent application for aiding COVID-19 diagnosis. MedRxiv. 2020.

34. Feng C, Huang Z, Wang L, Chen X, Zhai Y, Zhu F, et al. A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. Ann Transl Med. 2021;9(3):201.

35. Zhang XY, Huang HJ, Zhuang DL, Nasser MI, Yang MH, Zhu P, et al. Biological, clinical and epidemiological features of COVID-19, SARS and MERS and AutoDock simulation of ACE2. Infect Dis Poverty. 2020;9(1):1-1.

36. Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics. 2015;2015:198363.