ORIGINAL RESEARCH

Prediction of Disease Progression of COVID-19 Based upon Machine Learning

Fumin Xu,[1,]* Xiao Chen,[2,]* Xinru Yin,[1] Qiu Qiu,[3] Jingjing Xiao,[4] Liang Qiao,[5] Mi He,[5] Liang Tang,[6] Xiawei Li,[1] Qiao Zhang,[1] Yanling Lv,[1] Shili Xiao,[1] Rong Zhao,[1] Yan Guo,[1] Mingsheng Chen,[ID][5] Dongfeng Chen,[1] Liangzhi Wen,[1,7] Bin Wang,[1,8] Yongjian Nian,[5] Kaijun Liu[ID][1,8]

[1]Department of Gastroenterology, Daping Hospital, Army Medical University, Chongqing, People's Republic of China; [2]Department of Nuclear Medicine, Daping Hospital, Army Medical University, Chongqing, People's Republic of China; [3]Department of Gastroenterology, People's Hospital of Chongqing Hechuan, Chongqing, People's Republic of China; [4]Department of Medical Engineering, Xinqiao Hospital, Army Medical University, Chongqing, People's Republic of China; [5]College of Biomedical Engineering and Imaging Medicine, Army Medical University, Chongqing, People's Republic of China; [6]Department of Internal Medicine, 63790 Military Hospital of the Chinese People's Liberation Army, Xichang, People's Republic of China; [7]Department of Infectious Disease, Taikang Tongji Hospital, Wuhan, People's Republic of China; [8]Department of Infectious Disease, Wuhan Huoshenshan Hospital, Wuhan, People's Republic of China

*These authors contributed equally to this work

Correspondence: Kaijun Liu
Department of Gastroenterology, Daping Hospital, Army Medical University, 10 Changjiang Branch Road, Yuzhong District, Chongqing, 400038, People's Republic of China
Tel +86 135-9406-7813
Email kliu_tmmu@126.com

Background: Since December 2019, COVID-19 has spread throughout the world. Clinical outcomes of COVID-19 patients vary among infected individuals. Therefore, it is vital to identify patients at high risk of disease progression.

Methods: In this retrospective, multicenter cohort study, COVID-19 patients from Huoshenshan Hospital and Taikang Tongji Hospital (Wuhan, China) were included. Clinical features showing significant differences between the severe and nonsevere groups were screened out by univariate analysis. Then, these features were used to generate classifier models to predict whether a COVID-19 case would be severe or nonsevere based on machine learning. Two test sets of data from the two hospitals were gathered to evaluate the predictive performance of the models.

Results: A total of 455 patients were included, and 21 features showing significant differences between the severe and nonsevere groups were selected for the training and validation set. The optimal subset, with eleven features in the k–nearest neighbor model, obtained the highest area under the curve (AUC) value among the four models in the validation set. D-dimer, CRP, and age were the three most important features in the optimal-feature subsets. The highest AUC value was obtained using a support vector–machine model for a test set from Huoshenshan Hospital. Software for predicting disease progression based on machine learning was developed.

Conclusion: The predictive models were successfully established based on machine learning, and achieved satisfactory predictive performance of disease progression with optimal-feature subsets.

Keywords: COVID-19, disease progression, machine-learning models

Introduction

By November 22, 2020, more than 180 countries had reported a total of 57.8 million confirmed COVID-19 cases, a condition caused by SARS-CoV2.[1] SARS-CoV-2 is a novel enveloped RNA β-coronavirus, which has phylogenetic similarity to SARS-CoV, the pathogen causing SARS.[2] The clinical symptoms of COVID-19 have a broad spectrum, and vary among individuals.[3] Most infected individuals have mild or subclinical illness, while approximately 15.7%–32% of hospitalized COVID-19 patients develop severe acute respiratory distress or are admitted to an intensive care unit.[3,4] Potential risk factors to identify patients who will develop into severe or critical severe cases at an early stage include older age, underlying comorbidities, and elevated D-dimer.[5,6] As the COVID-19 outbreak continues to evolve, it is critical to find patients at high risk of disease progression. Several investigations have analyzed risk factors

associated with disease progression and clinical outcomes, and suggested that older age, comorbidities, immunoresponse were potential risk factors.[6–10] However, the clinical details were not well described, and many important laboratory results were not included in the analyses. Therefore, it is necessary to develop an effective classifier model for predicting disease progression at an early stage. Machine-learning techniques provide new methods for predicting clinical outcomes and assessing risk factors. Here, we aimed to predict the disease's progression with machine learning, based on a large set of clinical and laboratory features. Performance of the models was evaluated using clinical data of multicenter-confirmed COVID-19 patients. Software was developed for clinical practice. These predictive models can identify patients at high risk of disease progression and predict the prognosis of COVID-19 patients accurately.

Methods
Patients and Data Collection
This retrospective multicenter cohort study was performed at Huoshenshan Hospital and Taikang Tongji Hospital (Wuhan, China). Diagnostic criteria for COVID-19 followed the 2020 WHO interim guidance.[11] Severe COVID-19 cases were defined as patients with fever plus one of respiratory rate >30 breaths/minute, severe respiratory distress, or SpO_2 ≤93% in room air. All severe cases included had progressed from nonsevere cases. Adults with pneumonia but no signs of severe pneumonia and no need for supplemental oxygen were defined as nonsevere. All nonsevere cases study were stable and had been discharged. RT-PCR assays of nasal and pharyngeal swab specimens were performed for laboratory confirmation of SARS-CoV2 virus.

Data of COVID-19 patients were collected from February 10, 2020 to April 5, 2020. A total of 29 features of laboratory data obtained on admission to hospital (within 6 hours) are shown in Supplementary Table 1. This study was approved by the ethics committee of Huoshenshan Hospital (HSSLL024). As all subjects were anonymized in this retrospective study, written informed consent was waived due to urgent need. This study was conducted in accordance with the Declaration of Helsinki.

Study Design
A feature selection process was employed to incrementally choose the most representative features. The features with significant difference between two groups were selected for the following machine learning process. The combination training–validation set was collected from Huoshenshan Hospital, and two test sets were collected from Huoshenshan Hospital and TaikangTongji Hospital, respectively.

To prevent overfitting and improve generalizability, k-fold cross-validation was used. Since training and validation data were randomly generated, we took the average score of five rounds of k-fold cross-validation as the final validation results. The optimal-feature subsets in each model were defined as those with the highest AUC values. The flow diagram of training, validation, and test of the prediction models is shown in Figure 1.

Machine-Learning Models
Four prediction models were trained with logical regression (LR), support vector-machine(SVM), k–nearest neighbor (KNN), and naïve Bayes (NB), respectively. Experiments were implemented using MatLab 2018. ROC curve, AUC value, sensitivity, and specificity were used to evaluate predictive performance. The prediction tasks in this work mean classification.

Software for predicting disease progression of COVID-19 was developed based on machine learning, which is convenient for clinicians to use. The interface of the software is written in Visual Studio 2013 and the internal function in MatLab 2018.

Statistical Analysis
Statistical analyses were performed using SPSS 23.0. Categorical data are expressed as proportions. Descriptive data are expressed as medians and interquartile ranges for skewed-distribution variables and means ± SD for variables with normal distribution. Student's t-tests and nonparametric Mann–Whitney tests were used to compare normal- and skewed-distribution variables, respectively. Pearson's χ^2 was used to compare categorical variables and multiple rates. Two-sided $\alpha<0.05$ was considered statistically significant.
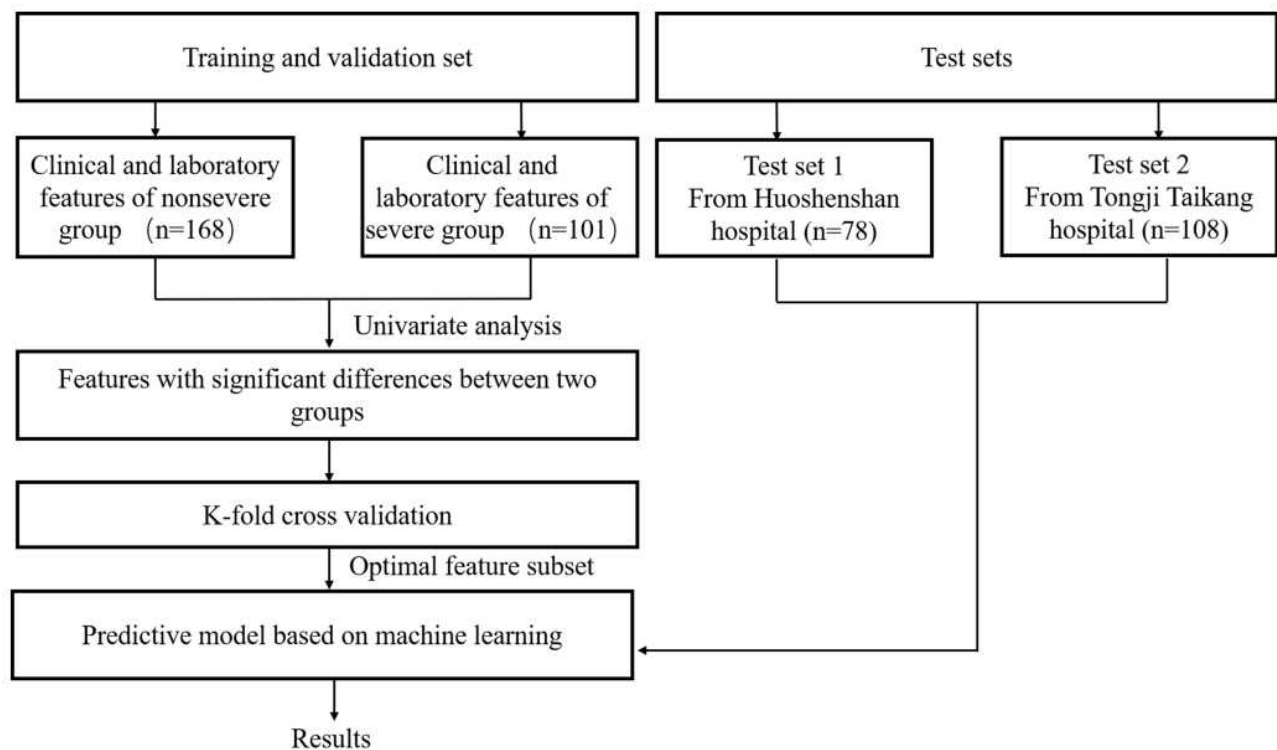
Figure 1 Flow diagram of training, validation, and testing of the prediction models.

Results

Demographic and Clinical Characteristics

By April 5, 2020, 1,567 COVID-19 patients in the medical record systems of Huoshenshan Hospital and Taikang Tongji Hospital had been screened for data collection. Data from 455 patients (347 from Huoshenshan, 108 from Taikang Tongji) with complete medical information and laboratory-examination results were collected. In sum, 78 patients from Huoshenshan were randomly selected as test set 1 (30 severe cases and 48 nonsevere cases) and 108 patients from Taikang Tongji as test set 2 (40 severe cases and 68 nonsevere cases). Data of the remaining 269 patients from Huoshenshan were used for the training and validation set (101 severe cases and 168 nonsevere cases). Demographic and clinical characteristics of the 269 patients in the training–validation set are summarized in Table 1, and clinical characteristics of patients in test sets 1 and 2 are summarized in Supplementary Table 2 and 3, respectively.

The median age of the patients in training and validation set was 62.75 years, and 51% of the patients were men. Severe patients were much older than non-severe patients (71.31 vs 57.61, $P<0.05$). Comorbidities were present in 55% of patients (147–270), and the prevalence of comorbidities in severe patients was higher than that in nonsevere patients (73% vs 45%, $P<0.05$). Hypertension (32%), diabetes (13%), and coronary heart disease (9%) were the most common comorbidities, and presented more frequently in severe patients: 26% of patients overall had two or more comorbidities, while severe patients had higher prevalence of two or more comorbidities (52% vs 15%, $P<0.05$). Fever (68%), cough (49.4%), and fatigue (45%) were the most common symptoms at onset of illness, and fever and fatigue were present more frequently in severe patients (Table 1).

Severe patients had elevated levels of CRP, lactate dehydrogenase (LDH), D-dimer, and α-hydroxybutyrate dehydrogenase, and had reduced levels of hemoglobin, hematocrit, and albumin (Table 1). Features with significant differences between the groups were introduced for selection using machine learning.

Predictive Performance of Machine-Learning Models in Validation Set

A total of 21 features with significant difference between the training and validation sets were used for the following modeling (Supplementary Table 4). The subset with the highest AUC was selected to be the

Table 1 Demographic and clinical characteristics of COVID–19 patients in training and validation sets

	All patients (n=269)	Nonsevere (n=168)	Severe (n=101)	*P*
Median age, years (IQR)	62.75 (61.08–64.39)	57.61 (55.64–59.64)	71.31 (69.02–73.5)	0
Male sex, n (%)	137 (51)	83 (49)	54 (53)	0.532
Onset time, days (IQR)	16.23 (15.05–17.51)	16.04 (14.77–17.32)	16.54 (14.05–19.04)	0.464
Comorbidity, n (%)	147 (55)	73 (45)	74 (73)	0
Hypertension, n (%)	87 (32)	36 (21)	51 (50)	0
Diabetes, n (%)	36 (13)	18 (11)	18 (18)	0.097
Coronary heart disease, n (%)	25 (9)	8 (4)	17 (17)	0.001
Chronic obstructive lung disease, n (%)	12 (4)	4 (2)	8 (8)	0.068
Carcinoma, n (%)	6 (2)	3 (2)	3 (3)	0.833
Chronic kidney disease, n (%)	5 (2)	1 (0.5)	4 (4)	0.13
Other, n (%)	32 (12)	13 (7)	19 (19)	0.007
Two or more, n (%)	69 (26)	26 (15)	43 (52)	0
Initial symptoms				
None, n (%)	8 (3)	4 (2)	4 (4)	0.713
Fever (temperature ≥37.3°C)	183 (68)	122 (73)	61 (60)	0.002
Cough, n (%)	133 (49.4)	90 (54)	43 (43)	0.081
Sputum, n (%)	16 (6)	7 (4)	9 (9)	0.111
Dyspnea, n (%)	56 (21)	27 (16)	29 (29)	0.013
Myalgia, n (%)	9 (3)	4 (2)	5 (5)	0.433
Fatigue, n (%)	120 (45)	53 (32)	67 (66)	0
Diarrhea, n (%)	6 (2)	3 (2)	3 (3)	0.833
Nausea or vomiting, n (%)	13 (5)	7 (4)	6 (6)	0.511
Complete blood-cell count				
White cells (×10⁹/L)	7.18 (6.62–7.80)	5.59 (5.34–5.58)	9.81 (8.46–11.16)	0
Neutrophils (×10⁹/L)	5.27 (4.66–5.84)	3.47 (3.28–3.69)	8.27 (6.91–9.62)	0
Lymphocytes (×10⁹/L)	1.35 (1.26–1.46)	1.59 (1.49–1.68)	0.97 (0.82–1.12)	0
Hemoglobin (g/L)	120.59 (118.21–122.83)	123.93 (121.08–126.79)	115.02 (110.93–119.11)	0
Hematocrit, %	36.11 (35.51–36.70)	36.94 (36.23–37.64)	34.73 (33.58–35.87)	0.001
Platelets (×10⁹/L)	236.48 (223.74–249.39)	262.21 (248.86–275.56)	193.67 (174.62–212.73)	0
Inflammatory markers				
CRP (mg/L)	28.18 (22.78–34.02)	5.68 (3.78–7.58)	65.60 (51.84–79.35)	0
Liver function				
Albumin (g/L)	34.99 (34.46–35.55)	36.82 (36.26–37.38)	31.93 (31.08–32.79)	0
ALT (IU/L)	35.21 (31.35–39.07)	34.27 (29.97–38.57)	37.96 (29.80–46.12)	0.697
AST (IU/L)	32.73 (25.85–39.61)	30.57 (20.66–40.48)	36.72 (29.58–43.86)	0.003
GGT (IU/L)	50.61 (43.33–57.89)	45.52 (36.39–54.64)	59.16 (47.09–71.24)	0.106
ALP (IU/L)	82.17 (76.65–89.95)	77.53 (69.19–85.86)	89.89 (81.75–98.04)	0.001
TBil (μmol/L)	11.76 (10.71–12.85)	9.96 (8.53–11.39)	14.75 (12.91–16.60)	0
DBil (μmol/L)	4.64 (3.85–5.44)	3.69 (2.64–4.73)	6.40 (5.30–7.49)	0
IBil (μmol/L)	6.99 (6.48–7.50)	6.02 (5.59–6.45)	8.76 (7.62–9.91)	0
Renal function				
BUN (mmol/L)	6.81 (5.54–9.28)	4.74 (4.46–5.02)	10.26 (5.49–15.03)	0
Creatinine (μmol/L)	75.61 (69.14–83.15)	70.28 (61.93–78.62)	84.50 (72.93–95.06)	0.018

(*Continued*)

Table 1 (Continued).

	All patients (n=269)	Nonsevere (n=168)	Severe (n=101)	P
Biochemical indices				
Glucose (mmol/L)	6.14 (5.80–6.51)	5.41 (5.13–5.69)	7.37 (6.66–8.09)	0
Calcium (mmol/L)	2.09 (2.08–2.11)	2.14 (2.13–2.16)	2.00 (1.98–2.03)	0
Myocardial enzymes				
LDH (IU/L)	270.10 (251.47–293.20)	198.92 (187.26–210.58)	388.50 (345.19–431.82)	0
Creatine kinase (IU/L)	95.50 (60.39–130.60)	65.09 (51.66–78.53)	150.22 (54.81–245.64)	0.060
Creatine phosphokinase–Mb (IU/L)	11.57 (10.17–12.96)	10.01 (8.69–11.33)	14.36 (11.29–17.43)	0
αHBDH (IU/L)	217.82 (202.19–234.64)	164.20 (155.14–173.26)	307.02 (270.52–343.52)	0
Coagulogram				
PT (s)	13.69 (13.17–14.21)	12.85 (12.67–13.02)	15.16 (13.81–16.51)	0
APTT (s)	30.31 (26.99–33.61)	28.07 (27.55–28.59)	34.19 (25.12–43.25)	0.214
INR	1.22 (1.11–1.34)	1.07 (1.05–1.08)	1.49 (1.18–1.80)	0
D-dimer (mg/L)	2.03 (1.58–2.50)	0.86 (0.63–1.09)	3.98 (2.86–5.10)	0

Notes: *P*-values calculated by *t*-test, Mann–Whitney *U*test, χ^2 test, or Fisher's exact test, as appropriate. Illness severity of COVID-19 was defined according to World Health Organization interim guidance.
Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT, γ-glutamyl transpeptidase; ALP, alkaline phosphatase; TBil, total bilirubin; BUN, blood urea nitrogen; LDH, lactate dehydrogenase; HBDH, hydroxybutyrate dehydrogenase; PT, prothrombin time; APTT, activated partial thromboplastin time; INR, international normalized ratio.

optimal subset of the corresponding machine-learning method (Table 2). Briefly, KNN achieved the highest AUC (0.9484, 95% CI 0.924–0.973) among the eleven features of the four methods in training and validation sets (Table 2). D-dimer was the single optimal feature with the highest AUC in the optimal-feature subset of each machine-learning method (0.8368 in LR model, 0.8169 in NB model, 0.8343 in KNN model, and 0.8322 in SVM model, respectively; Supplementary Table 5). ROC curves obtained by the optimal-feature subsets, single features, and all features using *k*-fold cross-validation are shown in Figure 2. Highest AUC values in optimal-feature subsets were 0.937, 95% CI 0.902–0.972) for LR, 0.949 (95% CI 0.924–0.973) for KNN, 0.935 (95% CI 0.906–0.964) for NB, and 0.931 (95% CI 0.895–0.967) for SVM (Table 3).

We compared predictive performance obtained by the models based on the optimal-feature subsets. Sensitivity (Sen), specificity (Spe), false-positive rate (FPR), false-negative rate (FNR), positive predictive value (PPV), negative predictive value (NPV), accuracy, and F1 scores of the above four models are shown in Table 3. No significant differences were observed among these four models for Sen, FNR, PPV, NPV, or accuracy. Spe, FPR, and F1 scores for SVM were superior (Table 3).

Predictive Performance of Each Feature of Optimal Subsets in Validation Set

To evaluate the importance of each feature in the corresponding optimal subsets, we evaluated predictive performance based on AUC obtained by each feature in the subsets. D-dimer, CRP, age, white blood cell (WBC) count, LDH, and albumin showed the highest predictive performance in the optimal subsets, with D-dimer, CRP, and age the top three (Supplementary Table 5).

Predictive Performance of Machine-Learning Models in Test Sets

Test set 1 comprised 78 patients from Huoshenshan, and test set 2 108 patients from Taikang Tongji. AUC values obtained by the four models in test set 1 were 0.9059 (95% CI 0.832–0.980) for LR, 0.9139 (95% CI 0.841–0.987) for KNN, 0.9177 (95% CI 0.848–0.988) for NB, and 0.9594 (95% CI 0.920–0.999) for SVM. F1 scores of the four models in test set 1 were 0.818 for LR, 0.828 for KNN, 0.867 for NB, and 0.885 for SVM (Supplementary Table 6). ROC curve obtained for the models in test set 1 are shown in Figure 3A. No significant differences were observed among these models for Sen, Spe, FPR, FNR, PPV, NPV, or accuracy (Supplementary Table 6). The predictive performance of all models was satisfied in test

Table 2 Optimal-feature subset of each machine learning method

	Age	Comorbidity	RR	PR	WBCs	Neutrophils	Lymphocytes	Hemoglobin	Hematocrit	Platelets	CRP	Albumin	ALP	TBil	BUN	Creatinine	Glucose	Calcium	LDH	αHBDH	D-dimer	AUC
LR	✓	✓	✓	✓				✓	✓	✓	✓	✓		✓							✓	0.9368
KNN	✓	✓	✓	✓	✓		✓				✓	✓									✓	0.9484
NB	✓		✓	✓	✓	✓		✓			✓	✓				✓				✓	✓	0.9349
SVM			✓	✓	✓									✓					✓		✓	0.9180

Abbreviations: RR, respiratory rate; PR, pulse rate; WBCs, white blood cells; HGB, hemoglobin; ALP, alkaline phosphatase; TBil, total bilirubin; BUN, blood urea nitrogen; LDH, lactate dehydrogenase; HBDH, hydroxybutyrate dehydrogenase.

set 1. Then, to test whether these models would still work at another hospital, we evaluated predictive performance in test set 2. AUC values of the four models in test set 2 were 0.8143 (95% CI 0.728–0.901) for LR, 0.8057 (95% CI 0.717–0.894) for KNN, 0.8265 (95% CI 0.741–0.912) for NB, and 0.8140 (95% CI 0.728–0.900) for SVM. F1 scores of the four models in test set 2 were 0.676 for LR, 0.698 for KNN, 0.716 for NB, and 0.691 for SVM (Supplementary Table 7). ROC curves obtained by the four models in test set 2 are shown in Figure 3 (Figure 3B). No significant differences were observed among these four models for Sen, Spe, FPR, FNR, PPV, NPV, or accuracy P>0.05, Supplementary Table 7).

To explore potential reasons for differences between the two test sets, we randomly selected 54 patients from test set 2 (Taikang Tongji), and added their data to the training–validation set. The remaining data from test sets 2 and 1 were combined (from Huoshenshan). As such, data from 323 patients were used as the training–validation set, and data from 132 patients were used as mixed test set. AUC values obtained by the four models were 0.8843 (95% CI 0.823–0.946) for LR, 0.8561 (95% CI 0.786–0.926) for KNN, 0.9096 (95% CI 0.853–967) for NB, and 0.9255 (95% CI 0.882–0.969) for SVM in the mixed test set. F1 score of the four models in the mixed test set were 0.777 for LR, 0.750 for KNN, 0.840 for NB, 0.832 for SVM, respectively (Supplementary Table 8). ROC curves obtained by the four models in test set 2 are shown in Figure 3C. The predictive performance of the models in the mixed test set was much better than that in test set 2 (Figure 3D).

Development of Software for Predictive Models

Software was developed for predicting disease progression based on machine learning for clinical practice (Supplementary Figure 1, 2, and 3). The first page provided the function of training and validation using *k*-fold cross-validation to select the optimal-feature subset and parameters (Supplementary Figure 1). In second page, one model that has been trained can be easily selected, and predictive performance can be evaluated in test sets (Supplementary Figure 2). Once the validity of the trained model has been confirmed by the second page, a prediction probability wil emerge for an upcoming patient on the third page (Supplementary Figure 3).
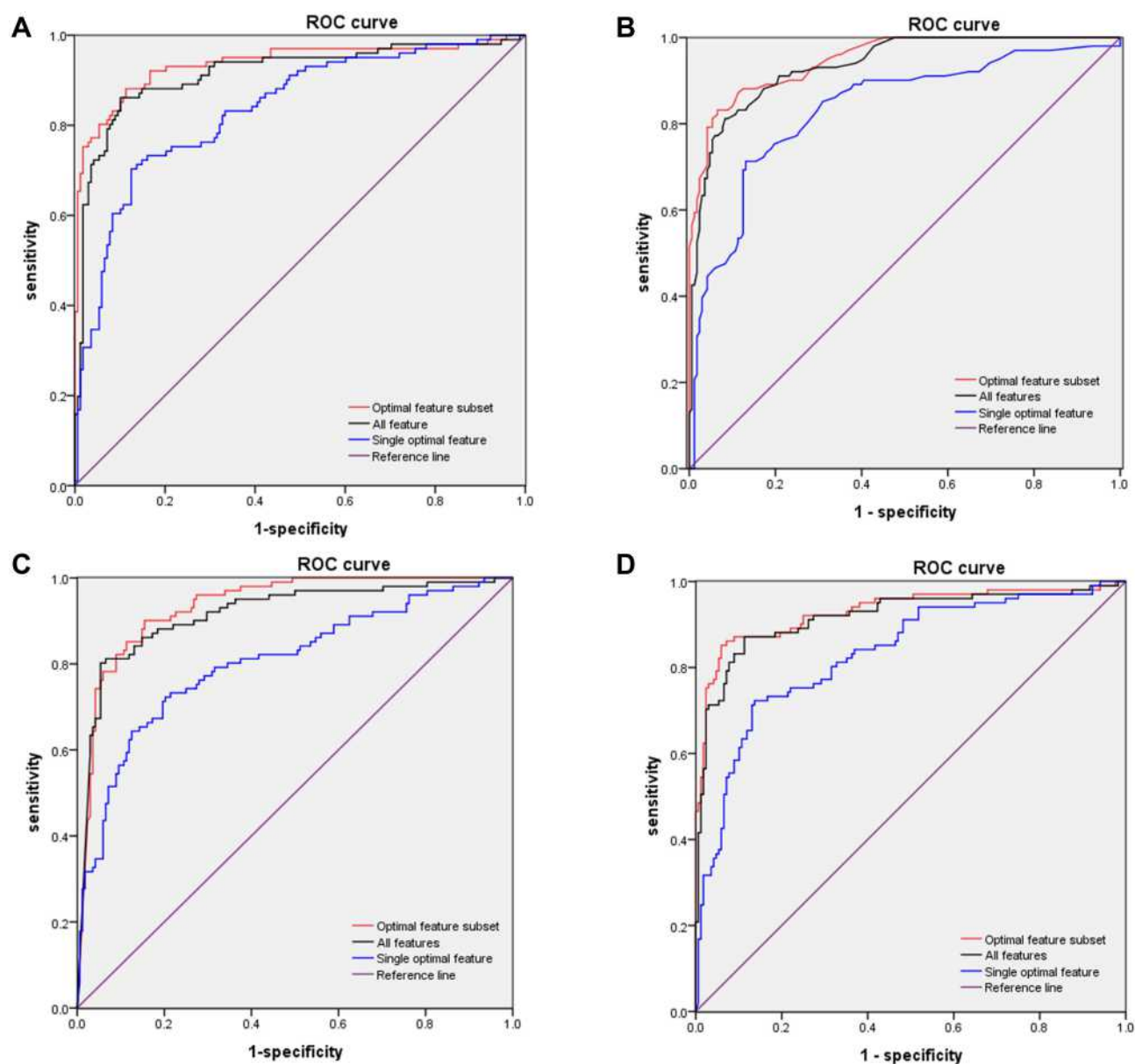
Figure 2 ROC curves for models in training and validation sets. (**A**) ROC curves of LR subsets for distinguishing between severe and nonsevere patients. AUC of optimal-feature subset 0.937 (95% CI 0.902–0.927), all features 0.916 (95% CI 0.876–0.955), and single optimal feature (D-dimer) 0.837 (95% CI 0.786–0.887). (**B**) ROC curves for subsets of features from KNN for distinguishing between severe and nonsevere patients. AUC of the optimal feature subset 0.948 (95% CI 0.924–0.937), all features 0.935 (95% CI 0.907–0.963), and single optimal feature (D-dimer) 0.835 (95% CI 0.782–0.887). (**C**) ROC curves of subsets of features from NB for distinguishing between severe and nonsevere patients. AUC of optimal feature set 0.935 (95% CI 0.906–0.964), all features 0.916 (95% CI 0.879–0.954), and single optimal feature (D-dimer) 0.805 (95% CI 0.748–0.861). (**D**) ROC curves of subsets of features from SVM for distinguishing between severe and nonsevere patients. AUC of optimal feature subset 0.931 (95% CI 0.895–0.967), features 0.918 (95% CI 0.879–0.957), and single optimal feature (D-dimer) 0.832 (95% CI 0.781–0.884).

Discussion

We developed a prediction model of disease progression based on machine learning. Clinical characteristics, WBC count, inflammatory markers, liver function, renal function, and coagulation functions were collected and utilized to establish the predictive model based on machine learning. In sum, 21 features with significant differences

between the severe and nonsevere groups were selected from a total of 48 features. In this feature-selection process, relatively useless features were eliminated to make the calculation more effective. Finally, the optimal-feature subset was determined using k-fold cross validation for each method. Moreover, the predictive performance of the models was evaluated by two test sets from two hospitals,

Table 3 Comparison of the average predictive performance by *k*-fold cross-validation with optimal-feature subset

	LR (95% CI)	KNN (95% CI)	NB (95% CI)	SVM (95% CI)	*P*
Sen	88.1% (81.7%–94.5%)	83.2% (75.7%–90.6%)	90.1% (84.2%–96.0%)	85.1% (78.1%–92.2%)	0.479
Spe	88.7% (83.9%–93.5%)	93.5% (89.7%–97.2%)[a]	84.5% (79.0%–90.0%)[b, c]	94.0% (90.4%–97.7%)[a]	0.01
FPR	11.3% (6.5%–16.1%)	6.5% (2.8%–10.3%)[a]	15.5% (10.0%–21.0%)[b, c]	6.0% (2.3%–9.6%)[a]	0.01
FNR	11.9% (5.5%–18.3%)	16.8% (9.4%–24.3%)	9.9% (4.0%–15.8%)	14.9% (7.8%–21.9%)	0.479
PPV	82.4% (75.1%–89.7%)	88.4% (81.9%–95.0%)[a]	77.8% (70.1%–85.4%)[b,c]	89.6% (83.4%–95.8%)[a]	0.065
NPV	92.5% (88.4%–96.6%)	90.2% (85.8%–94.7%)	93.4% (89.4%–97.4%)	91.3% (87.1%–95.6%)	0.736
Accuracy	88.5% (84.6%–92.3%)	89.6% (85.9%–93.3%)	86.6% (82.5%–90.7%)	90.7% (87.2%–94.2%)	0.48
F_1 score	0.8517	0.8517	0.8349	0.8731	—
AUC	0.937 (0.902–0.972)	0.949 (0.924–0.973)	0.935 (0.906–0.964)	0.931 (0.895–0.967)	—

Notes: [a]$P<0.05$ vs NB; [b]$P<0.05$ vs KNN; [c]$P<0.05$ vs SVM. *P*-values denote overall statistical results for the four models.

and AUC values in these test sets were satisfactory. We also developed software to predict disease progression of COVID-19 based on machine learning that can be used conveniently in clinical practice.

Clinical features of the patients in this study were consistent with previous large-sample studies.[3,12] Comorbidity, older age, lower lymphocyte count, and higher LDH were identified as independent high-risk factors for COVID-19 disease progression.[13] Ji et al developed a risk factor–scoring system (CALL) based on these features to predict disease progression.[13] However, there were few cases included, and the reliability of the model needs to be confirmed. In our study, these models were trained by optimal-feature subsets to attain optimal predictive performance. We evaluated predictive performance with two test sets from two hospitals to ensure the reliability of the models.

D-dimer, CRP, age, WBC count, LDH, and albumin had better predictive performance in the optimal-feature subset, with D-dimer, CRP, and age the top three. Zhou et al found no significant differences between a nonaggravation group and aggravation group for WBC count, CRP, albumin, LDH, or D-dimer level.[10] They found that total lymphocyte count was a risk factor associated with disease progression in COVID-19 patients using a binary logistic regression model.[10] However, only 17 patients were included in this study, and total lymphocyte count did not reflect disease progression. Zhou et al showed that older age and elevated D-dimer could help clinicians to identify patients with poor prognosis at an early stage.[6] Consistently with this study, age and D-dimer level were

important features in the optimal-feature subset. Elevated levels of D-dimer are associated with disease activity and inflammation, mainly including venous thromboembolism, sepsis, or cancer.[14,15] A retrospective study on deceased patients also showed that D-dimer was markedly higher in deceased patients than recovered patients.[16] Therefore, monitoring the D-dimer levels can help clinicians identify patients at high risk of disease progression. Anticoagulation treatment can be given patients with high D-dimer levels to prevent disease progression. Albumin levels decrease significantly in most severe COVID-19 patients and decrease continuously during the disease's progress.[17] Hypoalbuminemia is associated with poor clinical outcomes for hospitalized patients.[18,19] Hypoalbuminemia in severe patients is mainly due to inadequate nutrition intake and overconsumption.

The predictive performance of the models in test set 1 was much better than that in test set 2. and patients enrolled in test set 2 were from another hospital. Differences in laboratory findings and medical services may be potential reasons for the lower predictive performance in test set 2. After data from Taikang Tongji had been added to this training set, predictive performance improved significantly, indicating that predictive performance in another hospital could be improved if part of the data collected from another hospital participated in the training stage.

Code Availability

The code of the software used in this study is available from the corresponding author on reasonable request.
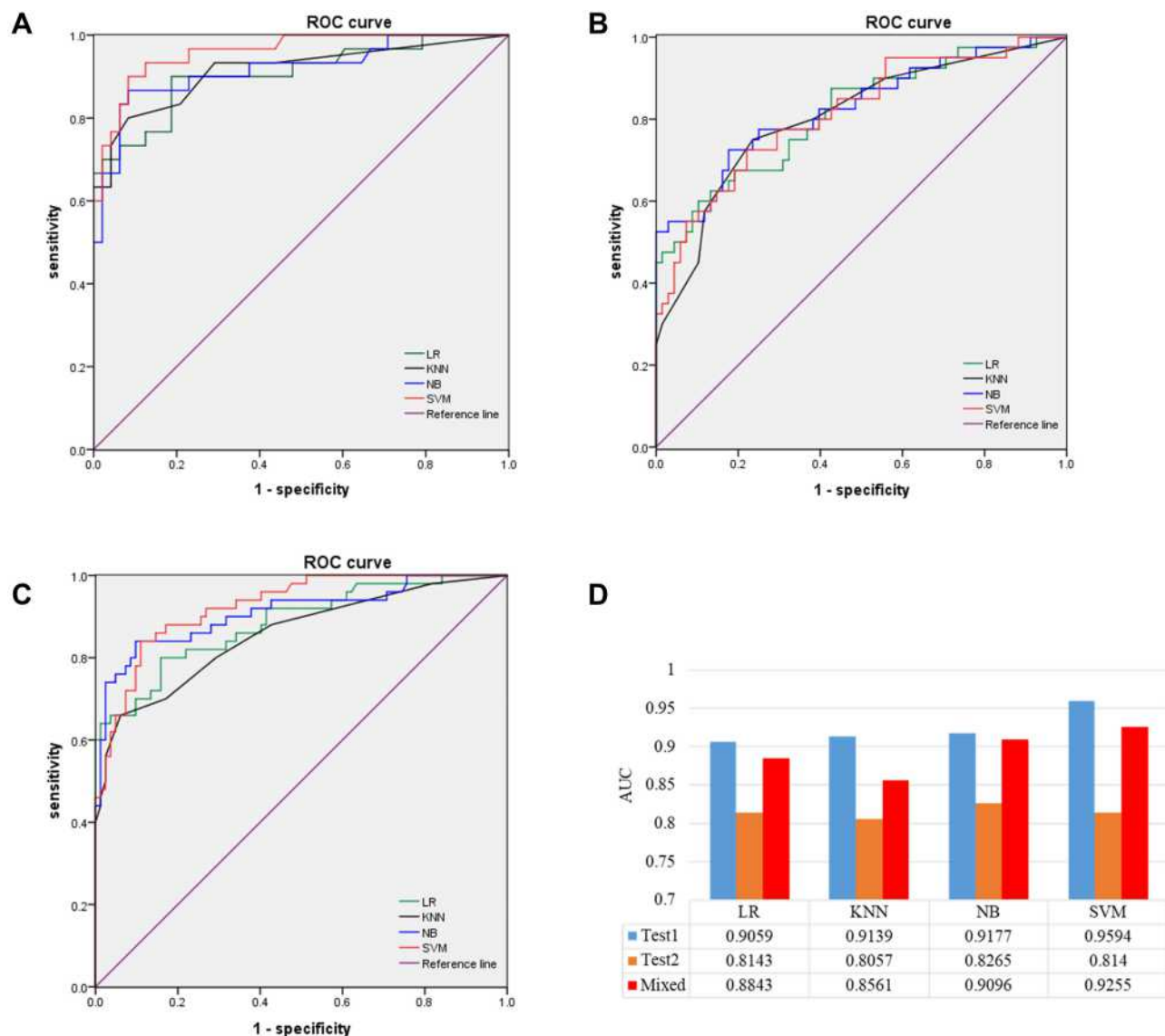
Figure 3 ROC curves of models in testing sets. (**A**) Optimal-feature set of LR, KNN, NB, and SVM in test set 1. (**B**) Optimal feature set of LR, KNN, NB, and SVM in test 2. (**C**) Optimal-feature set of LR, KNN, NB, and SVM in the mixed test sets. (**D**) AUC values of optimal-feature subsets for different models in test set 1, test set 2, and mixed test set.

Data Sharing Statement

The data sets used in this study are available from the corresponding author — Kaijun Liu (email kliu_tmmu@126.com) on reasonable request.

Ethics Approval

This study was approved by the ethics committee of Huoshenshan Hospital (Wuhan, China) (HSSLL024).

Informed Consent

As all subjects were anonymized in this retrospective study, written informed consent was waived due to urgent need.

Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data, took part in drafting the article or revising it critically for important intellectual content, agreed to submit to the current journal, gave final approval to the version to be published, and agree to be accountable for all aspects of the work.

Funding

(cstc2017jcyjAX0302, cstc2020jcyj-msxmX1100), and Army Medical University Frontier Technology Research Program (2019XLC3051). The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the study, and had final responsibility for the decision to submit for publication.

Disclosure

The authors declare that there are no conflicts of interest.

References

1. World Health Organization. Weekly epidemiological update - 24 November 2020. Available from: https://www.who.int/publications/m/item/weekly-epidemiological-update—24-november-2020. Accessed April 07, 2021.
2. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565–574. doi:10.1016/S0140-6736(20)30251-8
3. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497–506. doi:10.1016/S0140-6736(20)30183-5
4. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223):507–513. doi:10.1016/S0140-6736(20)30211-7
5. Sun Y, Koh V, Marimuthu K, et al. Epidemiological and clinical predictors of COVID-19. *Clin Infect Dis*. 2020;71(15):786–792. doi:10.1093/cid/ciaa322
6. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395(10229):1054–1062. doi:10.1016/S0140-6736(20)30566-3
7. Guan WJ, Liang WH, Zhao Y, et al. Comorbidity and its impact on 1590 patients with Covid-19 in China: a nationwide analysis. *Eur Respir J*. 2020;55(5):2000547. doi:10.1183/13993003.00547-2020
8. Wang L, He W, Yu X, et al. Coronavirus Disease 2019 in elderly patients: characteristics and prognostic factors based on 4-week follow-up. *J Infect*. 2020;80(6):639–645. doi:10.1016/j.jinf.2020.03.019
9. Wu C, Chen X, Cai Y, et al. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. *JAMA Intern Med*. 2020;180(7):934. doi:10.1001/jamainternmed.2020.0994
10. Zhou Y, Zhang Z, Tian J, Xiong S. Risk factors associated with disease progression in a cohort of patients infected with the 2019 novel coronavirus. *Ann Palliat Med*. 2020. doi:10.21037/apm.2020.03.26
11. World Health Organziation. Clinical management of severe acute respiratory infection when Novel coronavirus (nCoV) infection is suspected: interim guidance. 2020. Available from: https://www.who.int/docs/default-source/coronaviruse/clinical-management-of-novel-cov.pdf. Accessed April 7, 2021.
12. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708–1720. doi:10.1056/NEJMoa2002032
13. Ji D, Zhang D, Xu J, et al. Prediction for progression risk in patients with COVID-19 pneumonia: the CALL score. *Clin Infect Dis*. 2020;71(6):1393–1399. doi:10.1093/cid/ciaa414
14. Borowiec A, Dabrowski R, Kowalik I, et al. Elevated levels of d-dimer are associated with inflammation and disease activity rather than risk of venous thromboembolism in patients with granulomatosis with polyangiitis in long term observation. *Adv Med Sci*. 2020;65(1):97–101. doi:10.1016/j.advms.2019.12.007
15. Schutte T, Thijs A, Smulders YM. Never ignore extremely elevated D-dimer levels: they are specific for serious illness. *Neth J Med*. 2016;74(10):443–448.
16. Chen T, Wu D, Chen H, et al. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. *BMJ*. 2020;368:m1091. doi:10.1136/bmj.m1091
17. Zhang Y, Zheng L, Liu L, Zhao M, Xiao J, Zhao Q. Liver impairment in COVID-19 patients: a retrospective analysis of 115 cases from a single center in Wuhan city, China. *Liver Int*. 2020;40(9):2095–2103. doi:10.1111/liv.14455
18. Kim S, McClave SA, Martindale RG, Miller KR, Hurt RT. Hypoalbuminemia and clinical outcomes: What is the mechanism behind the relationship? *Am Surg*. 2017;83(11):1220–1227. doi:10.1177/000313481708301123
19. Yanagisawa S, Miki K, Yasuda N, Hirai T, Suzuki N, Tanaka T. Clinical outcomes and prognostic factor for acute heart failure in nonagenarians: impact of hypoalbuminemia on mortality. *Int J Cardiol*. 2010;145(3):574–576. doi:10.1016/j.ijcard.2010.05.061