The Classifiers Established with Clinical Laboratory Indicators to Distinguish COVID-19 from Community-Acquired Pneumonia: Retrospective Cohort Study

Wanfa Dai, Pei-Feng Ke, Zhen-Zhen Li, Qi-Zhen Zhuang, Wei Huang, Yi Wang, Yujuan Xiong, Xian-Zhang Huang

Submitted to: Journal of Medical Internet Research
on: August 11, 2020

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

The Classifiers Established with Clinical Laboratory Indicators to Distinguish COVID-19 from Community-Acquired Pneumonia: Retrospective Cohort Study

Wanfa Dai[1*]; Pei-Feng Ke[2*]; Zhen-Zhen Li[3]; Qi-Zhen Zhuang[3]; Wei Huang[1]; Yi Wang[2]; Yujuan Xiong[2*] PhD; Xian-Zhang Huang[2*] MD

[1]GongAn County People's Hospital Gong An CN
[2]the Second Affiliated Hospital of Guangzhou University of Chinese Medicine Guangzhou CN
[3]Second Clinical Medical College, Guangzhou University of Chinese Medicine Guangzhou CN
[*]these authors contributed equally

Corresponding Author:
Yujuan Xiong PhD
the Second Affiliated Hospital of Guangzhou University of Chinese Medicine
111 Dade Rd., Guangzhou
Guangzhou
CN

Abstract

Background: The initial symptoms of the patients with COVID-19 are very much alike with those of the patients with community-acquired pneumonia (CAPN), and it is difficult to distinguish COVID-19 from CAPN by clinical symptoms and imaging examination.

Objective: The objective of our study was to construct an effective model for early identification of COVID-19 from CAPN.

Methods: The clinical laboratory indicators (CLIs) of 61 COVID-19 patients and 60 CAPN patients were analyzed retrospectively. Random combinations of various CLIs (CLI_combinations) were utilized to establish COVID19_vs_CAPN classifiers with machine learning algorithms including Random Forest Classifier (RFC), Logistic Regression (LR) and Gradient Boosting Classifier (GBC). The performance of the classifiers was assessed using the area under the receiver operating characteristic curve (AUC) and recall rate in COVID-19 prediction with the test data.

Results: The classifiers constructed with three algorithms from 43 CLI_combinations showed high performance (recall rate > 0.9 and AUC > 0.85) in COVID-19 prediction for the test_set. In the high performance classifiers, the CLIs including PCT (procalcitonin), MCHC (mean corpuscular hemoglobin concentration), UA (urine acid), albumin, AGR (ratio of albumin to globulin), NEUTC (neutrophil count), RBC (red blood cell count), monocyte coun, BASOC (basophil count) and WBC (white blood count) showed a high usage rate, they also had high feature_importance except BASOC. The CLI_combination of [PCT, AGR, UA, WBC, NEUTC, BASOC, RBC, MCHC] was the representative one of nine optimal CLI_combinations capable of constructing perfect classifiers (AUC = 1.0) with RFC or GBC, and replacing any CLI in these CLI_combinations would lead to a significant degradation in the performance of classifiers built with them.

Conclusions: The classifiers constructed with only a few specific CLIs could perfectly distinguish COVID-19 from CAPN, which will help clinicians with early isolation and centralized management of COVID-19 patients.

(JMIR Preprints 11/08/2020:23390)
DOI: https://doi.org/10.2196/preprints.23390

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).

✓ **Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will**

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

The Classifiers Established with Clinical Laboratory Indicators to Distinguish COVID-19 from Community-Acquired Pneumonia: Retrospective Cohort Study

Wanfa Dai[2#],Pei-Feng Ke[1#], Zhen-Zhen Li[3], Qi-Zhen Zhuang[3], Wei Huang[2], Yi Wang[1], Yujuan Xiong[1*], Xian-Zhang Huang[1*]

#Contributed equally

Author Affiliations:

[1] Department of Laboratory Medicine, the Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangdong Provincial Key Laboratory of Research on Emergency in TCM, Guangzhou, 510120, China;

[2] Department of Respiration, GongAn County People’s Hospital, Gong An, Hubei Province, 434300, P.R.China.

[3] Second Clinical Medical College, Guangzhou University of Chinese Medicine, Guangzhou, 510120, China;

* **corresponding author:**

Xian-Zhang Huang M.D, Department of Laboratory Medicine, the Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangdong Provincial Key Laboratory of Research on Emergency in TCM, 111 Dade Rd., 510120, Guangzhou, P.R. China, E-mail: huangxz020@gzucm.edu.cn;

Yujuan Xiong Ph.D, Department of Laboratory Medicine, the Second Affiliated Hospital of Guangzhou University of Chinese Medicine, 111 Dade Rd., 510120, Guangzhou, P.R. China, E-mail: yujuanxiong@gzucm.edu.cn.

Abstract

Background: The initial symptoms of the patients with COVID-19 are very much alike with those of the patients with community-acquired pneumonia (CAP), and it is difficult to distinguish COVID-19 from CAP by clinical symptoms and imaging examination.

Objective: The objective of our study was to construct an effective model for the early identification of COVID-19 from CAP.

Methods: The clinical laboratory indicators (CLIs) of 61 COVID-19 patients and 60 CAP patients were analyzed retrospectively. Random combinations of various CLIs (CLI_combinations) were utilized to establish COVID19_vs_CAP classifiers with machine learning algorithms including Random Forest Classifier (RFC), Logistic Regression (LR) and Gradient Boosting Classifier (GBC). The performance of the classifiers was assessed using the area under the receiver operating characteristic curve (AUC) and recall rate in COVID-19 prediction with the test data.

Results: The classifiers constructed with three algorithms from 43 CLI_combinations showed high performance (recall rate > 0.9 and AUC > 0.85) in COVID-19 prediction for the test_set. In the high-performance classifiers, the CLIs including procalcitonin (PCT), mean corpuscular hemoglobin concentration (MCHC), urine acid (UA), albumin, ratio of albumin to globulin (AGR), neutrophil count (NEUTC), red blood cell count (RBC), monocyte count, basophil count (BASOC) and white blood count (WBC) showed a high usage rate. they also had high feature importance except BASOC. The feature combination (FC) of [PCT, AGR, UA, WBC, NEUTC, BASOC, RBC, MCHC] was the representative one among the nine FCs used to constructed the classifiers with an AUC equal to 1.0 by using RFC or GBC. Replacing any CLI in these FCs would lead to a significant reduction in the performance of the classifiers built with them.

Conclusions: The classifiers constructed with only a few specific CLIs could efficiently distinguish COVID-19 from CAP, which would help clinicians perform the early isolation and the centralized management of COVID-19 patients.

Keywords: COVID-19; clinical laboratory indicators; community-acquired pneumonia; classifier; classification algorithm

Introduction

The coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) infection, that was discovered in early December 2019, has become a global pandemic. As of 3 August, 2020, COVID-19 has become widespread in 215 countries, areas, or territories worldwide; it has caused infection in more than 17.9 million people, and has resulted in the deaths of more than 686,000 people . The World Health Organization has stated that the COVID-19 spread may be impeded by early detection, isolation, and implementation of a robust healthcare system . Nevertheless, the published data indicated that the initial symptoms of COVID-19 patients are very similar to those of the patients with common cold or influenza. The COVID-19 patients have different clinical symptoms, and some of them do not have any symptoms . SARS-COV-2 infection has a long incubation period, with a median incubation period of 5–7 days, which is the chief risk factor for community infection . Community-acquired pneumonia (CAP) and COVID-19 have similar clinical and imaging features, but their treatment and infectivity are very dissimilar. Isolating COVID-19 from CAP is very important to prevent the spread of COVID-19 and to provide the specific treatment.

Some characteristic spectra demonstrated by clinical laboratory indicators (CLI) of COVID-19 patients have been utilized as auxiliary clues for diagnosis . Previous studies have demonstrated that increased procalcitonin (PCT), lymphocytopenia, and thrombin activation can all be utilized as auxiliary diagnostic indicators of COVID-19 and poor prognostic factors . However, they are also correlated with CAP . Thus, in accordance with the changes of these indicators, it is impossible to differentiate COVID-19 from CAP. The changes of neutrophil/lymphocyte ratio and peak platelet/lymphocyte ratio, lactate dehydrogenase (LDH), C-reactive protein (CRP), and interleukin-6 (IL-6) are considered to be associated with the progression and prognosis of COVID-19 , but using the information of the CLIs to give clinicians correct guidance is still a great challenge.

Classifiers established by machine learning (ML) algorithms based on various clinical features, biomarkers, and CLIs are increasingly widely utilized in disease diagnosis and risk prediction . In the

COVID-19 pandemic, ML was also widely used to predict, classify, assess, track, and control the spread of SARS-COV-2 . ML can improve diagnostic performance compared with hand-selected biomarkers by selecting relevant biomarkers and more consistently capturing both their relative importance to prediction and their interactions among one another . In this study, we used CLIs to build classifiers with different ML algorithms to distinguish COVID-19 patients from CAP patients, and found that only the feature combinations (FCs) with many specific CLIs rather than the FCs with the most significantly differential CLIs between the two groups could build high-performance classifiers (HPC).

Materials and methods

Collection of patient’s electronic medical record data

The electronic medical records of patients who were admitted in Gong’an County People’s Hospital and diagnosed with COVID-19 or CAP from December 2019 to March 2020 were retrieved. The information regarding the patient’s age, sex, clinical symptoms upon admission, medical history, epidemiological history, computed tomography (CT) imaging features, and CLIs were sorted out for retrospective analysis. Only the laboratory test results during admission were included. It was specified that the data of all patients were kept confidential, and the data of all patients were only utilized for comprehensive analysis. No personal information of any patient was mentioned in the paper. This study was approved by the ethics committees from the Guangdong Provincial Hospital of Chinese Medicine (approval number: ZE2020-027-01) with a waiver of informed consent due to the retrospective nature of the study.

Data description

Diagnosis and clinical classification of COVID-19 were performed according to “Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment (7th edition)”. Sixty-one patients with COVID-19 and 60 patients with CAP were enrolled according to the discharge diagnosis on the electronic medical record. There were 3 mild, 47 common, 6 severe and 5 critical types, which were categorized into two groups in further analysis, as follows: COVID19-COM (3 mild and 47 common type) and COVID19-SV (6 severe and 5 critical types). They were age and sex matched, and did not significantly differ in terms of medical history. The main clinical symptoms between CAP and COVID-19 groups were not significantly different.

Primary analysis

The descriptive analysis of all CLIs was performed between groups or sub-groups. Between-group or between-subgroup differences were tested using the python module ‘statsmodels’ . Student’s t-test was performed when the distribution of the variables conforms to the normal distribution. Otherwise, Mann–Whitney U test was used. Chi-square test was used to detect the difference of counting data in baseline data between two groups or subgroups. A value of $P < 0.05$ was considered to be significant.

Feature selection and data pre-processing

The CLIs with a missing value ratio greater than 20% were directly excluded. Only the CLIs with a significant difference between the two groups were selected and used to generate 1807780 non-repetitive random FCs (consisting of one to eight CLIs) by using the iterator ‘combinations’ in the Python module ‘itertools’ . Next, a FC was selected from the FC list one by one to form a new data sheet with the dependent variable (disease type), and 1807780 new data sheets were eventually formed. For each new data sheet, the rows with missing values were removed. Then the remained rows were divided into training_dataset and test_dataset using Scikit-learn (v.0.23.1) (function ‘train_test_split’ with test_size = 0.25, random_state = 0). The training_dataset was used to build the classifier, and the test dataset was used to assess the performance. The feature values were standardized using ‘StandardScaler’ function in Scikit-learn module before constructing the logistic regression classifier.

Construction of classifiers with ML algorithms in Scikit-learn module

Scikit-learn is a Python module integrating a wide range of state-of-the-art ML algorithms for medium-scale supervised and unsupervised problems . Logistic regression classifier (LR), random forest classifier (RFC) and gradient boosting classifier (GBC) were typically used to construct classifiers in prediction of disease risk, progression, prognosis, and so on . LR classifier in the ‘sklearn.linear_model’ is also known as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function . RFC in the ‘sklearn.ensemble’ module is one of the averaging algorithms in ensemble methods and is a perturb-and-combine techniques specifically designed for trees. In random forests. Each tree in the ensemble is built from a sample drawn with replacement from the training_set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size setting with parameter ‘max_features’. In practice, the variance reduction due to the introduction of randomness in the classifier construction is often significant hence yielding an overall better model . GBC (using ‘sklearn.ensemble’ function) is a boosting method, in which base estimators are built sequentially. To reduce the bias of the combined estimator, one has to combine several weak models to produce a powerful ensemble. GBC builds an additive model in a forward stage-wise fashion, and it allows for the optimization of arbitrary differentiable loss functions .

In this study, the classifiers were respectively constructed using LR, RFC and GBC in Scikit-learn module with the training_dataset. The model parameter settings were kept as default except that the random_state was modified to ‘0’ for all models, and class_weight was modified to ‘balanced’ for LR and RFC models. The performance of the classifiers was evaluated with the test dataset by calculating the recall rate (sensitivity), specificity, accuracy, and area under the receiver operating characteristic curve (AUC) (using the ‘sklearn_metrics.recall_score’, ‘sklearn_metrics.precision_score’, ‘sklearn_metrics.accuracy_score’ and ‘sklearn_metrics.auc’ function respectively). Gini importance was computed (using the ‘feature_importances’ function) to measure the importance of each feature in the RFC and GBC classifier. The higher the Gini importance value, the more important the feature All the above analyses were performed in the python environment (version 3.7).

Results

Basic characteristics of CAP group and COVID-19 group

No significant difference in age and sex was found between CAP and COVID-19 groups (Table 1), but the proportion of males in both CAP and COVID-19 groups was 55% and 65.57%, higher than that of females, respectively. No significant difference in the medical history between the two groups (Table 1) was observed. Also, no significant difference was found in the proportion of main clinical symptoms between the two groups, such as fever, cough, fatigue, muscle soreness, and loss of appetite (Table 1). The average hospitalization days for CAP patients were remarkably lower than those for COVID-19 patients (*P* < .001). In the CAP group, some patients with pulmonary CT also had imaging features including patchy hyperdense shadow (18.33%), ground glass shadow (6.67%), and fibrous focus (10%). Nonetheless, the chief imaging features of pulmonary CT in the COVID-19 group were patchy hyperdense shadow (40.98%) and ground glass shadow (14.75%), and many patients (11.48%) had both patchy hyperdense shadow and ground glass shadow (Table 1). Among the 61 patients suffering from COVID-19, 3 (4.92%) had mild symptoms, 47 (77.05%) had common symptoms, 6 (9.84%) had severe symptoms, and 5 (8.20%) had critical symptoms. Fever and cough were the principal symptoms in the early stage of COVID-19, and these accounted for 70.49% and 63.93% of the cases, respectively (Table 1). Among the CAP patients included in the analysis, no cases of death were found during hospitalization, three of the five severely ill patients in the COVID-19 group aged 36, 49, and 74 years old died during hospitalization. The 36-year-old patient who died underwent interventricular septal repair in childhood.

Table 1. Comparison of baseline information between COVID-19 patients and CAP patients.

	CAP[a](n=60)	COVID-19[b](n=61)	*P*_values
Sex (male %)	33 (55%)	40 (65.57%)	0.27
Age (mean ± SD)	55.72±18.10	50.23±16.95	0.09
Hospitalization days (*median, 1st quartile, 3rd quartile*)	9 (7,12)	21(13,26)	< 0.001
Medical history:			
hypertension	14(23.33	16(26.23%)	0.83

	%)		
Diabetes	2 (3.33%)	6 (9.84%)	0.27
Liver disease	2 (3.33%)	3 (4.92%)	0.99
Heart Disease	3 (5.0%)	5 (8.20%)	0.72
exposure history	unclear	54 (88.52%)	-
Familial aggregation infection[c]	unclear	22 (36.07%)	-
The initial symptoms:			
fever	36 (60%)	43 (70.49%)	0.26
cough	44 (73.33%)	39 (63.93%)	0.33
Myalgia	4 (6.67%)	7 (11.48%)	0.53
poor appetite	5 (8.33%)	11 (18.03%)	0.18
fatigue	33 (55%)	24 (39.34%)	0.10
Time from onset of symptoms to admission (days)[*median, 1st quartile, 3rd quartile*]	unrecorded	3(1,7)	-
Imaging features:			
Patchy high-density opacity	11 (18.33%)	25 (40.98%)	0.009
Ground-glass opacity	4(6.67%)	9 (14.75%)	0.24
fibrotic lesion	6 (10%)	3 (4.92%)	0.32
Patchy high-density opacity and ground-glass opacity	0	7 (11.48%)	0.01
Death cases	0	3(4.92%)	-

Note: a. CAP: Patients with Community-Acquired Pneumonia; b. Patients with COVID-19 infection; c. There are more than 2 cases of infection after aggregation with family members or relatives.

Characteristic profile of the CLIs in COVID-19 and CAP

Even though most CLIs had a similar variation trend in both CAP and COVID-19, the extent of change was different. Among more than 60 evaluated CLIs, there were significant differences in 25 CLIs between the two groups (Table 2). A decrease of lymphocyte (LYM), red blood cell count (RBC), hematocrit (PCV), hemoglobin concentration (HGB) and mean corpuscular hemoglobin concentration (MCHC) and an increase of neutrophil ratio (NEUT), prothrombin time (PT), trace C-reactive protein (mCRP), and PCT were observed in both COVID-19 and CAP. Furthermore, the level of NEUT, PT, mCRP, and PCT in CAP was remarkably higher than those in COVID-19. Levels of LYM, RBC, PCV, HGB, and MCHC in CAP were significantly lower than those in COVID-19 (Figure 1). Various erythrocyte-related CLIs, RBC, PCV, HGB, and MCHC significantly decreased in both CAP and COVID-19 but with a greater reduction in CAP patients (Figure 1). The standard deviation of red blood cell distribution width (RDW-SD) and mean red blood cell volume (MCV) also indicated prominent differences between CAP and COVID-19 (Figure 1).

Table 2. Difference in Laboratory indicators between patients with CAP and COVID-19.

CLIs	CAP(n=60)			COVID-19(n=61)			*P*_val
	count[a]	mean	std[b]	count[a]	mean	std[b]	
PCT (ng/ml)	43	0.629	0.838	55	0.134	0.184	< 0.001
MAO-B (U/L)	35	4.569	1.748	53	3.538	1.592	0.001
MYO (ng/ml)	14	39.179	29.421	23	65.794	87.039	0.04
mCRP (mg/L)	41	63.943	64.530	13	22.568	29.577	0.004
PT (sec)	30	12.780	0.873	53	12.460	1.107	0.04
TT (sec)	30	15.123	1.565	53	14.655	1.422	0.049
ALB (g/L)	53	35.508	5.929	54	37.831	6.169	0.04
AGR(ALB/GLB)	53	1.211	0.295	54	1.378	0.482	0.047
AFU (U/L)	35	17.709	5.167	50	22.106	5.698	< 0.001
UA (μmol/L)	44	284.193	118.608	54	325.261	92.914	0.007
K (mmol/L)	54	3.900	0.462	55	4.021	0.392	0.03
WBC (×10^9/L)	58	8.858	5.576	56	5.293	2.047	< 0.001
NEUT (%)	57	72.958	15.544	56	66.661	14.013	0.007
LYM (%)	56	18.646	13.416	56	24.014	11.175	0.002
NEUTC(×10^9/L)	56	6.797	5.525	56	3.649	1.949	< 0.001
MOC (×10^9/L)	55	0.565	0.337	56	0.404	0.194	0.009
EOC (×10^9/L)	55	0.111	0.213	56	0.053	0.072	0.03
BASOC (×10^9/L)	55	0.021	0.013	56	0.015	0.013	0.002
RBC (×10^{12}/L)	56	4.028	0.647	56	4.284	0.570	0.008
HGB (g/L)	55	120.800	17.326	56	130.143	16.888	0.005
PCV (L/L)	55	0.371	0.052	56	0.389	0.049	0.04
MCV (fl)	55	93.255	6.662	56	91.241	6.501	0.01
MCHC(g/L)	55	325.473	8.360	56	334.482	13.559	< 0.001
RDW-SD (fl)	55	41.476	2.573	56	41.141	4.082	0.01

Note: a, the sample number after ruling out the miss values. b, std: standard deviation. MOC: monocyte count; EOC: Eosinophil count; BASOC: Basophil percentage. UA: urine acid; K: potassium;

Comparing between COVID19-COM and COVID19-SV, 26 CLIs demonstrated a remarkable difference (Table 3). In comparison with COVID19-COM, LDH, aspartate aminotransferase (AST), fibrinogen content (FIB), mCRP, and erythrocyte sedimentation rate (ESR) increased acutely in COVID19-SV, whereas prealbumin (PA), carbon dioxide binding capacity (CO_2CP), LYM, and lymphocyte count (LYMPH) decreased in COVID19-SV (Supplemental Figure 1).

An orderly increase of fucosidase (AFU), myoglobin (MYO), uric acid (UA), and MCHC and an orderly decrease of thrombin time (TT), monocyte count (MOC), eosinophil count (EOC), MCV, and RDW-SD were observed in CAP, COVID19-COM, and COVID19-SV patients, indicating that these

CLIs may be used to distinguish CAP from COVID-19 and may suggest the probability of severe COVID-19 progression (Supplemental Figure 2).

Table 3. Difference in laboratory indicators between common and severe patients with COVID-19

	Common type(n=50)			**Severe type(n=11)**			***P*_val**
CLIs	**count[a]**	**mean**	**std[b]**	**count[a]**	**mean**	**std[b]**	
PCT (ng/ml)	44	0.112045	0.169933	11	0.223636	0.21736	0.01
Pro-BNP (pg/ml)	29	366.0531	549.429	11	534.7818	398.0666	0.03
SCRP (mg/L)	41	23.33171	34.48338	11	72.45818	60.8048	0.002
LDH (U/L)	26	214.8962	73.3192	8	314.75	118.7552	0.02
D-dimer (mg/L)	42	0.833571	1.115034	11	5.132727	10.39914	0.005
MYO (ng/ml)	16	49.22125	60.50511	7	103.6743	127.354	0.02
cTn (ng/ml)	16	0.010625	0.0025	7	0.032857	0.041115	0.02
CK (U/L)	27	81.2963	47.15344	8	202.125	195.052	0.02
FIB (mg/dl)	42	411.9048	104.3626	11	467.4545	76.50015	0.03
AST (U/L)	46	29.41304	15.75588	10	45.6	18.96898	0.004
γ-GGT (U/L)	44	46.04545	41.60945	10	80	44.2292	0.007
ALB (g/L)	44	38.60227	6.266837	10	34.44	4.557826	0.02
AGR（ALB/GLB）	44	1.436364	0.506741	10	1.12	0.229976	0.02
IBIL (µmol/L)	44	9.481818	3.840741	10	7.96	4.335948	0.048
PA (mg/L)	41	180.1707	83.37383	9	125.5556	68.18195	0.03
β2-MG (mg/L)	41	1.977561	0.430121	9	2.527778	1.014702	0.01
CO_2CP (mmol/L)	41	25.41951	2.53685	9	22.73333	2.018044	0.002
K (mmol/L)	44	4.056591	0.414134	11	3.876364	0.250969	0.04
ESR (mm/h)	30	55.43333	41.63941	7	87	35.08086	0.02
NEUT (%)	45	64.49556	13.28586	11	75.51818	14.0007	0.02
LYM (%)	45	25.71111	10.93153	11	17.07273	9.750496	0.01
EO (%)	45	1.235556	1.388124	11	0.390909	1.037742	0.009
EOC ($\times10^9$/L)	45	0.062	0.075607	11	0.013636	0.0388	0.003
LYMPH ($\times10^9$/L)	45	1.255111	0.558137	11	0.834545	0.38258	0.008
PCV (L/L)	45	0.394511	0.049988	11	0.367545	0.036269	0.03
RDW-CV (%)	45	12.65778	1.170759	11	12.87273	0.781141	0.03

Note: a, the sample number after ruling out the miss values. b, std: standard deviation. Pro-BNP: N-terminal pro-B-type natriuretic peptide; SCRP: Hypersensitive C-reactive protein; MYO: Myoglobin; cTn: Troponin; CK: creatine kinase; γ-GGT: transglutaminase transpeptidase gamma;

IBIL: Indirect bilirubin; PA: prealbumin; β2-MG: β 2-microglobulin; K: potassium; ESR: erythrocyte sedimentation rate; EO: Eosinophil percentage; EOC: Eosinophil count; RDW-CV: Coefficient of variation of erythrocyte distribution width.

Classifiers constructed from the FCs with 7 to 8 CLIs could accurately distinguish COVID-19 from CAP

The performance of the classifiers was gradually improved as the number of CLIs in the FCs increased from one to eight. However, when the number of CLIs in the FCs reaches to eight, the performance of the classifiers constructed by these FCs was no longer significantly improved. The performance of the LR classifiers constructed by the FCs with 8 CLIs (8_CLI_combination) was even slightly lower than those constructed by the FCs with 7 CLIs (7_CLI_combination). Forty-three FCs, including five 7_CLI_combinations and 38 8_CLI_combinations, were determined according to that the recall rate. The AUCs of the classifiers constructed with LR, RFC and GBC were respectively greater than 0.9 and 0.85 (Supplemental Table 1). The ROC and PRC of the classifiers constructed with RFC, LR and GBC from the representative Senven_CLI_combination [PCT, AGR, UA, NEUTC, BASOC, MCV, MCHC] showed very high performance and precision in COVID-19 prediction, and the AUCs were 1.0, 0.97 and 0.96, respectively (Figure 2A), and the average precision were respectively 1.0, 0.97 and 0.98 (Figure 2B). The ROC of the classifiers constructed with RFC, LR and GBC from the representative Eight_CLI_combination [PCT, ALB, UA, WBC, MOC, BASOC, RBC, MCHC] showed AUCs of 1.0, 0.90 and 1.0, respectively (Figure 2C). The ROC of the classifiers constructed with the three algorithms from the Senven_CLI_combination [AGR, AFU, LYM, NEUTC, EOC, MCV, MCHC] showed AUCs of 0.98, 0.91 and 0.97, respectively (Figure 2D). Feature_importance results showed that BASOC was the least important in the above two representative CLI_combinations, and AFU was the most important in the CLI_combination (Figure 3). However, when BASOC was substituted with AFU in the above mentioned two CLI_combinations, the performance of the classifiers constructed with the new CLI_combinations decreased (Figure 2E and 2F). PCT and AFU were not observed to be in a same CLI_combination from which a HPC could be constructed. The above evidence and the fact that only

43 FCs with seven CLIs or 8 CLIs_could be used to build HPCs suggested that only the FCs with specific CLIs can establish HPCs to distinguish COVID-19 from CAP.

The importance of different CLIs in classifiers varied greatly, and the importance of the same CLI varied greatly among classifiers constructed by different FCs (Figure 3). In the HPCs constructed with the 7_CLI_combinations, the average feature importance of AFU (26.60%) was the highest, followed by UA (25.31%) and PCT (21.06%) (Figure 3A). However, in the HPCs constructed with the 8_CLI_combinations, the average feature importance of UA (22.51%) was the highest, followed by PCT (20.88%) and MCHC (12.36%) (Figure 3B). PCT and MCHC were very important to each classifier because they were respectively included in 100% (38/38) and 92.11% (35/38) of the 8_CLI_combinations (Figure 3B) and in 40% (2/5) and 100% (5/5) of the 7_CLI_combinations (Figure 3A). UA was also included in all 8_CLI_combinations, but its feature importance varied from 11.3% to 41.2% in different classifiers (Figure 3B).

Discussion

Principal Findings

The main highlight of this study is that only a few of the common CLIs were required to establish the classifier models to accurately distinguish COVID-19 from CAP. The HPCs could only be constructed by combining several specific CLIs. Among near two million FCs with 1-8 CLIs, only 43 FCs could be used to construct HPCs with a recall rate greater than 0.9 and an AUC greater than 0.85 to distinguish COVID-19 from CAP.

Comparison with Prior Work

We have established many high-performance COVID-19_Vs_CAP classifiers with FCs consisting of only CLIs, and almost no similar researches on distinguishing COVID-19 from CAP were reported. However, many studies have used CLIs to build ML models to help with COVID-19 diagnosis. The prediction performance of these models varied; the accuracy of these models in predicting COVID-19 was between 0.8 and 0.96 [36-38]. In addition, most of the reported ML models for the diagnosis or prediction of COVID-19 have been involved in more types of variables, such as CT results, clinical symptoms and CLIs [17, 38, 39]. Although most of these COVID-19-related ML models were built with more than two ML algorithms, but not all models constructed with each algorithm showed high performance. The methods of feature selection used in these studies included the recursive feature elimination algorithm [37], causal explanation models[17], and the least absolute shrinkage and selection operator (LASSO) regression [38]. These methods can extract the features that are closely related to the target phenotype, but whether the classifier constructed by the combination of these features has the best performance needs to be determined. The optimized FCs in this study were selected by evaluating the recall rate and AUC for each FC with 1-8 randomly selected CLIs from the differential CLIs between COVID-19 and CAP groups and by constructing classifiers using each FC with LR algorithm. The preliminarily screened FCs were used to build classifiers with RFC and GBC algorithms, and finally only the FCs capable of building the HPC simultaneous with LR, RFC and GBC algorithms were selected for the final model construction.

Limitations

As reported earlier, many inflammatory factors, including IL-6 and IL-10, are closely related to COVID-19 and have diagnostic value, but both IL-6 and IL-10 were not detected in the patients of this study. Cristina Menni et al. [18] reported that loss of smell and taste is a strong predictor for COVID-19. Deviations and omissions may exist in the patients' self-reported clinical symptoms. Thus, we did not take into account the clinical symptoms when building the classifiers. The possibility that other indicators are more important in constructing COVID-19_vs_CAP classifiers was not ruled out. In addition, the sample size included in this study is relatively small, and the classifiers needs to be optimized with larger samples before it can be used to distinguish COVID-19 from CAP in practice.

The rationality of the research results

Forty FCs contain PCT and MCHC among the 43 FCs. The feature importance of PCT in each classifier is very high, suggesting that PCT may be an good blood marker to efficiently distinguish COVID-19 from CAP. PCT is one of the markers of lower respiratory tract bacteria and other infections. The United States Food and Drug Administration approved the monitoring of the beginning and the entire duration of antibiotic treatment for suspected lower respiratory tract infections based on serum PCT levels . However, the elevation of serum PCT in COVID-19 patients was also reported in many studies . The increase of PCT is a remarkable characteristics of patients with COVID-19 . Increased serum PCT level in both COVID-19 and CAP patients indicated that suggestions for the identification of COVID-19 from CAP could not be made simply on the basis of the increase of PCT level. Compared with the normal reference value of the CLIs, the serum levels of the most of CLIs increased or decreased simultaneously in both COVID-19 and CAP patients. Thus, providing references for the diagnosis of COVID-19 or CAP directly according to the rise or decrease of the CLIs is difficult. However, However, we found that the ML classifiers constructed with the FCs with many certain CLIs could distinguish COVID-19 from CAP effectively, suggesting an advantage of ML algorithm in disease classification or diagnosis. .

The COVID-19_vs_CAP classifiers with the highest performance were also involved in PCT, MCHC, UA, ALB, NEUTC, MOC, BASOC, RBC and WBC, proposing the importance of these CLIs in differentiating COVID-19 from CAP. Few studies have reported the changing trend of MCHC in patients with COVID-19 or CAP, but the results of this study showed that MCHC decreased in both groups, and significantly lower in the CAP group than in the COVID-19 group. The reason for the decrease of MCHC may be closely related to the reduction of iron due to inflammation . The 1st-3rd quartiles of UA in both COVID-19 and CAP groups were within the normal reference range, but it was significantly higher in the COVID-19 group than in the CAP group. Elevated UA is an independent risk factor of renal injury or renal dysfunction; the underlying mechanisms of UA elevation are very complicated . The significant difference in UA between COVID-19 and CAP may be interpreted as follows: individuals with higher UA may be more susceptible to COVID-19 than those with lower UA levels. UA exists in all the 8_CLI combinations capable of constructing HP_CLFs, and has a high feature importance in the classifiers, suggesting that UA is another important marker that can distinguish COVID-19 from CAP. Zhou et al reported that ALB significantly decreased in severe and critical COVID-19 patients . Serum ALB level is a good prognostic marker in CAP. Decreased ALB level is closely associated with a higher risk of mortality in patients with CAP . Although ALB decreased remarkably in both COVID-19 and CAP groups, there was still a significant difference between the two groups, and the decrease in the CAP group was more obvious than that in the COVID-19 group, which could contribute to the differentiation of COVID-19 from CAP. AFU contributed high feature_importance in the HPCs constructed from 7_CLI_combinations due to the significant difference in AFU between COVID-19 and CAP. An increase of serum AFU has a certain diagnostic value for primary liver cancer. Thus the higher AFU in the COVID-19 group than that in the CAP group may be explained by the fact that liver injury is more common in COVID-19 than in CAP, or the diversity in AFU levels determines the difference in susceptibility to COVID-19.

Recommendations

Although both PCT and AFU contributed high feature_importance in the HPCs constructed from the FCs containing PCT or AFU, but the performance of the classifiers constructed from the FCs containing both PCT and AFU decreased remarkably. This result indicated that intrinsic dependence exist among some CLIs that undergo synergistic changes in individuals and can be used to construct HPCs. The internal relationship between CLIs is very complex and difficult to deconstructed. Therefore, that the following method may be effective: random selection of different CLIs to construct classifiers with different classification algorithms, followed by the evaluation of the performance of each classifier, and finally, the discovery of the FCs with certain CLIs that can be used to accurately distinguish COVID-19 from CAP.

Conclusions

The patients suffering from COVID-19 and CAP have their own characteristic profile of CLI, and some FCs consisting of 7 or 8 specific CLIs could build high-performance COVID-19_Vs_CAP classifiers. The usage rate and the feature_importance of the CLIs in the HPCs indicated that PCT, MCHC, UA, ALB, AGR, NEUTC, RBC, MOC and WBC, are the most important indicators that can distinguish COVID-19 from CAP.

Note

Acknowledgments: We are very grateful to Associate Professor Benrong Liu of the Second Affiliated Hospital of Guangzhou Medical University for his guidance in data analysis. We are also very grateful to the doctors and nurses of Gong’an County People’s Hospital for their hard work to prevent as well as control COVID-19.

Financial support: None.

Conflicts of interest: All authors declared that there were no conflicts of interest.

Author contributions: HXZ and XY made substantial contributions to the study concept and design. XY and KPF were in charge of the manuscript draft. WB and HW were responsible for obtaining ethical approval, collecting data, and confirming the data accuracy. LZZ, ZQZ and WY Sharifu did the analysis and interpretation. All authors agreed on the final version for submission to the journal.

Abbreviations

AFU: fucosidase

AGR: ratio of albumin to globulin

ALB: albumin

AST: aspartate aminotransferase

AUC: area under the curve

BASOC: basophil count

CAP: Community-acquired pneumonia

CLI: clinical laboratory indicators

CO2CP: carbon dioxide binding capacity

COVID-19: The coronavirus disease 2019

CRP: C-reactive protein

CT: computed tomography

EOC: Eosinophil count

ESR: erythrocyte sedimentation rate

FC: feature combination

FIB: fibrinogen content

GBC: Gradient Boosting Classifier algorithm

HGB: hemoglobin concentration

HPC: high performance classifier

IL-6: interleukin-6

LDH: lactate dehydrogenase

LR: Logistic Regression algorithm

LYM: lymphocyte

LYMPH: lymphocyte count

MaxEnt: maximum-entropy classification

MCHC: mean corpuscular hemoglobin concentration

mCRP: trace CRP

MCV: mean red blood cell volume

ML: machine learning

MOC: monocyte count

MYO: myoglobin

NEUT: neutrophil ratio

NEUTC: neutrophil count

PA: prealbumin

PCT: procalcitonin

PCV: hematocrit

PRC: precision-recall curve

PT: prothrombin time

RBC: red blood cell count

RDW-SD: red blood cell distribution width

RFC: Random Forest Classifier algorithm

ROC: receiver operating characteristic curve

SARS-COV-2: severe acute respiratory syndrome coronavirus 2

TT: thrombin time

UA: uric acid

WBC: white blood count

References

Figure legends

Figure 1 The statistical distribution of the plasma level of the CLIs with a remarkable difference between COVID-19 and CAP.

The statistical distribution was presented with a box and whisker plot. The horizontal lines within the boxes indicate the median value. The vertical lines extending below and above the boxes represent 5%–95% percentile values. The scale on the Y axis represents the values of the 5th, 25th, 50th, 75th, and 95th percentiles of the clinical laboratory index in the CAP group. The triangle represents the upper and lower limits of the normal reference range of the laboratory index, respectively.

Figure 2 ROC and precision recall curve plotted for the COVID-19_vs_CAP classifiers built with various FCs of different CLIs. At the top of each image is the CLI combination for constructing classifiers using three different classification algorithms.

Figure 3 Usage rate and the feature_importance of each CLI in the high-performance COVID-19_vs_CAP classifiers. A) The mean feature_importance of each CLI in the HPCs constructed with the 7_CLI_combinations; B) The mean feature_importance of each CLI in the HPCs constructed with the 8_CLI_combinations. The histogram is represented by an mean ± standard deviation. The number of the shadow backgrounds represents the minimum and maximum values of the feature_importance of CLI. The number indicated by the triangle symbol indicates the mean feature_importance of CLI in all classifiers. The number indicated by the circle indicates the usage rate of CLI in HPC. The number in the parentheses indicates how many CLI combinations are capable of constructing the HPCs containing the CLI.

Supplemental Figure 1 The statistical distribution of the plasma level of the CLIs with a significant difference between COVID19-COM and COVID19-SV.

The statistical distribution was presented with a box and whisker plot. The horizontal lines within the boxes indicate the median value. The vertical lines extending below and above the boxes represent 5%–95% percentile values. The scale on the Y axis represents the 5th, 25th, 50th, 75th, and 95th percentile values of the clinical laboratory index in the COVID19-COM subgroup. The triangle represents the upper and lower limits of the normal reference range of the laboratory index. The median of the CLI in COVID19-SV is also represented in the Y axis.

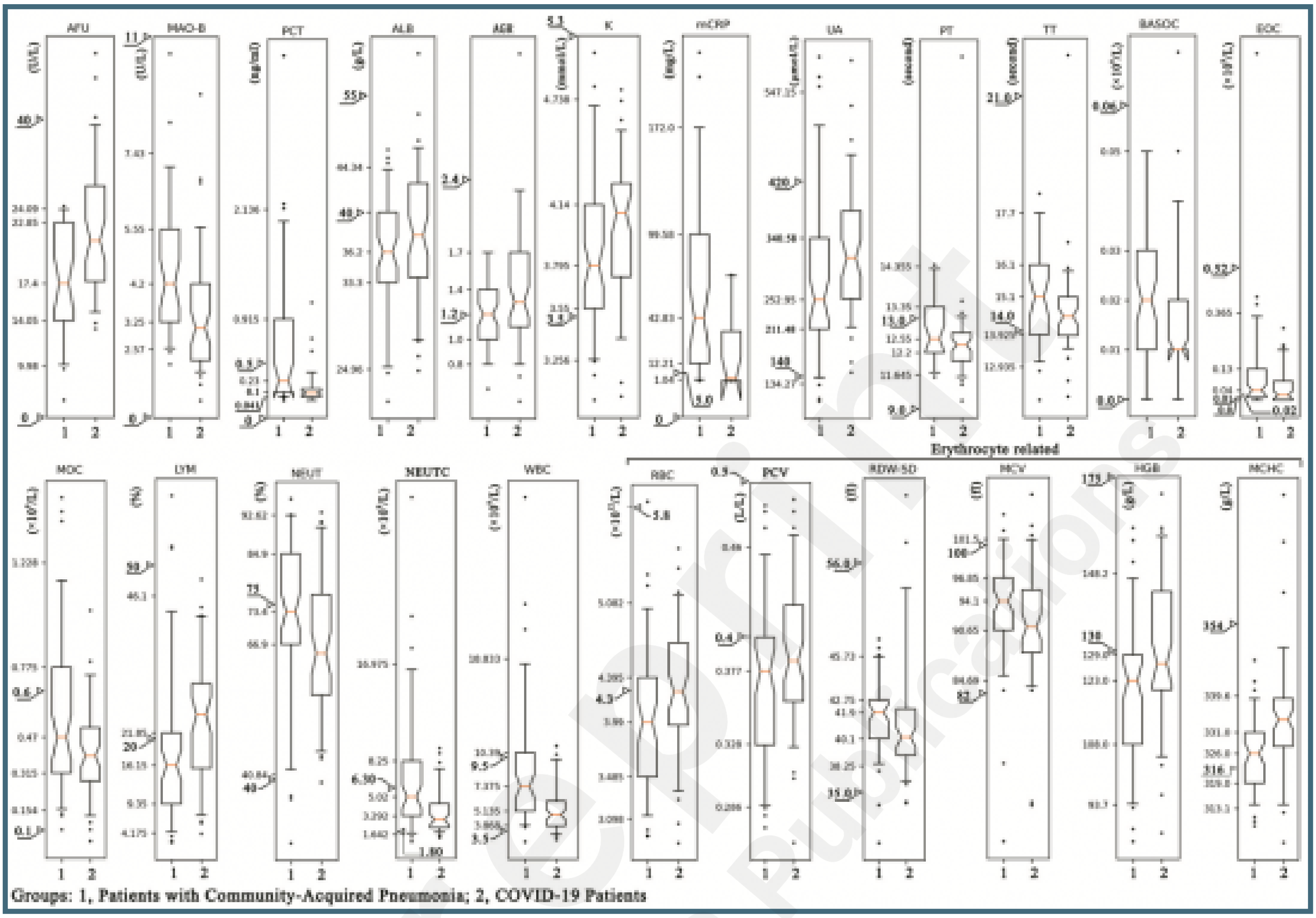
Supplemental Figure 2 The statistical distribution of the plasma level of the CLIs among CAP, COVID19-COM, and COVID19-SV.

The statistical distribution was presented with a box and whisker plot. The horizontal lines within the boxes indicate the median value. The vertical lines extending below and above the boxes represent 5%–95% percentile values. The scale on the Y axis represents the 5th, 25th, 50th, 75th, and 95th percentile values of the clinical laboratory index in the CAP group. The triangle represents the upper and lower limits of the normal reference range of the laboratory index.

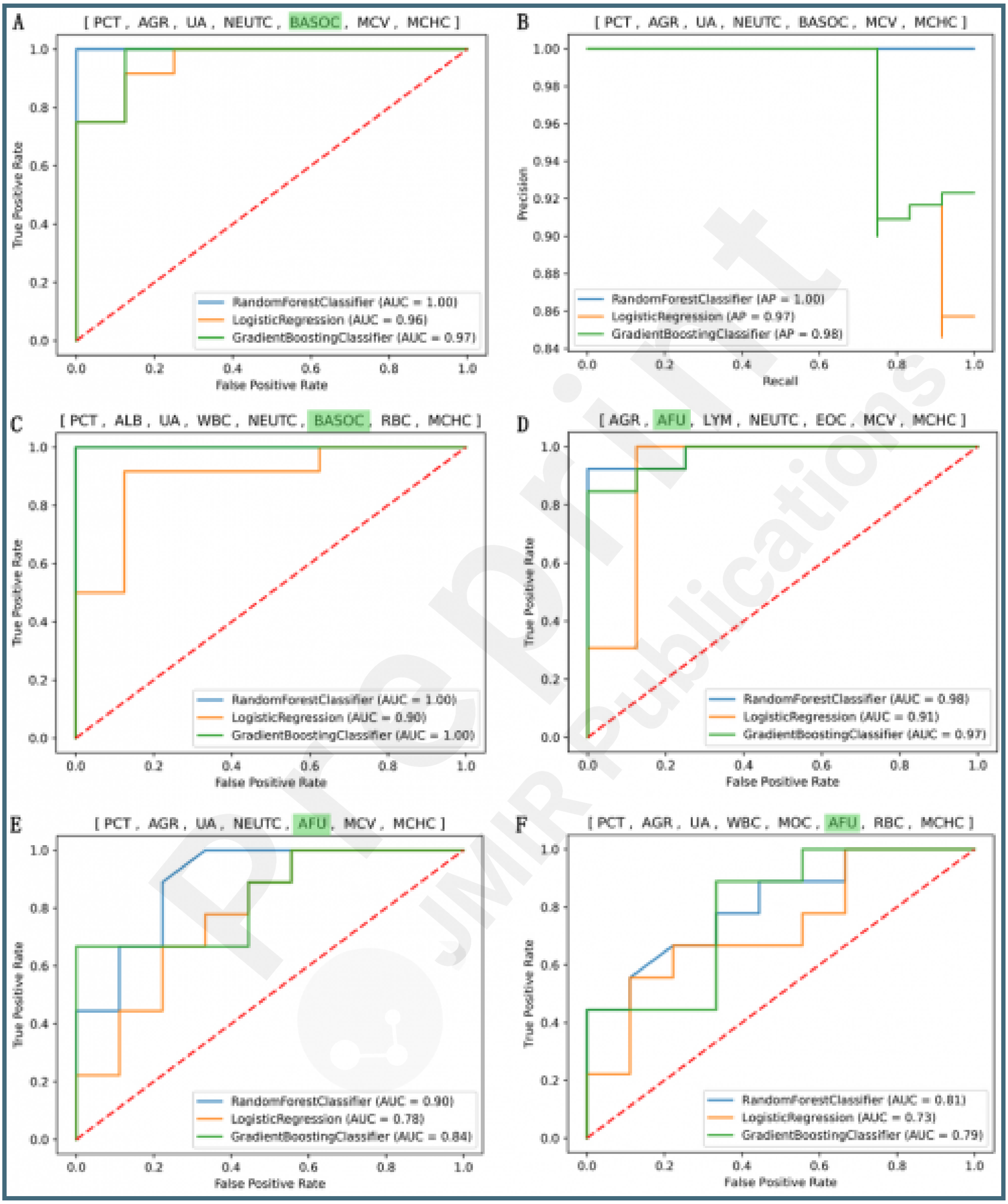
Supplementary Files

Figures

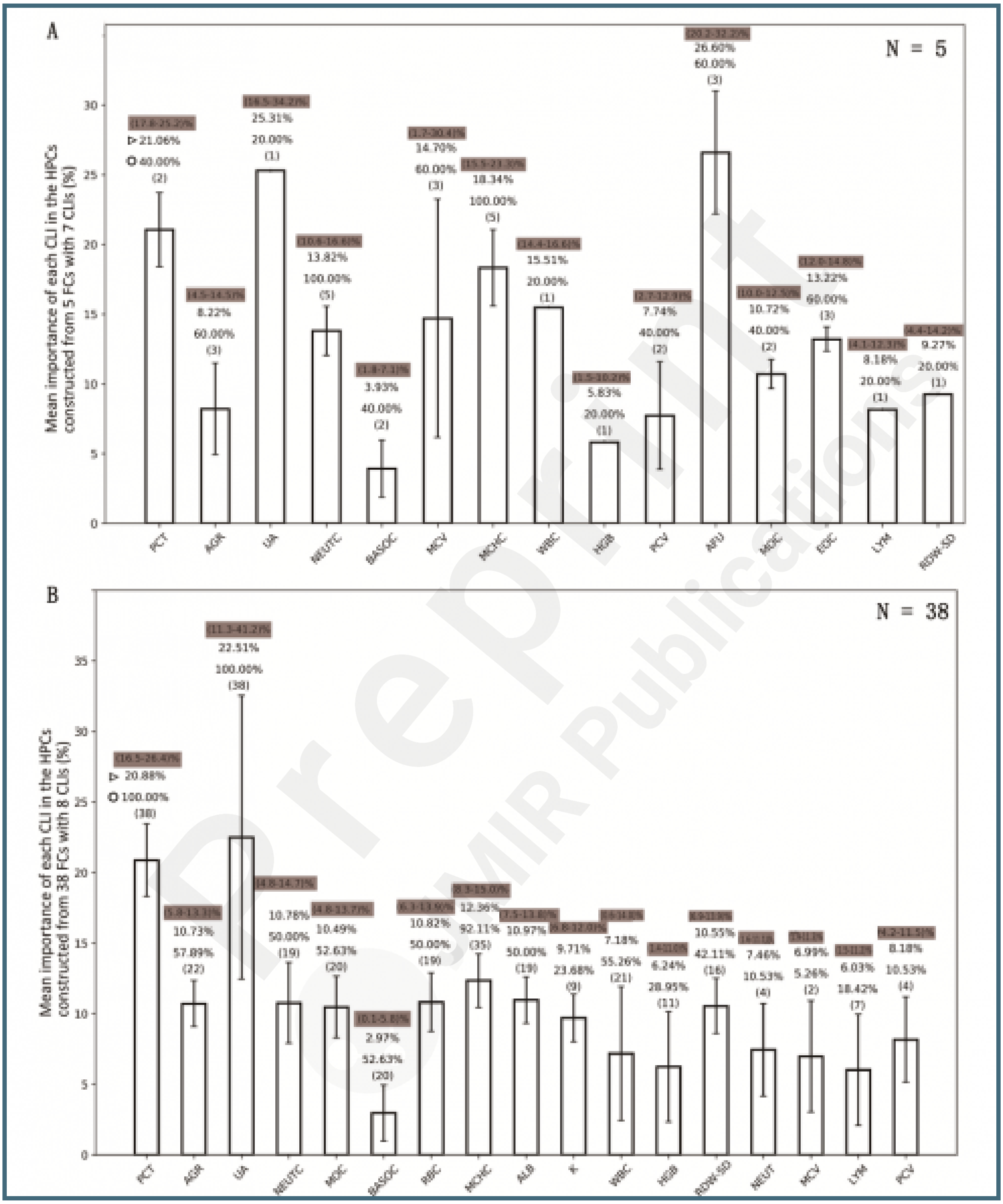
The statistical distribution of the plasma level of the CLIs with a remarkable difference between COVID-19 and CAP.



ROC and precision recall curve plotted for the COVID-19_vs_CAP classifiers built with various FCs of different CLIs.



Usage rate and the feature_importance of each CLI in the high-performance COVID-19_vs_CAP classifiers.



Multimedia Appendixes

Supplemental Figure 1 The statistical distribution of the plasma level of the CLIs with a significant difference between COVID19-COM and COVID19-SV.
URL: https://asset.jmir.pub/assets/b73fc55115fdf21406ea1553cfa1aa0d.png

Supplemental Figure 2 The statistical distribution of the plasma level of the CLIs among CAP, COVID19-COM, and COVID19-SV.
URL: https://asset.jmir.pub/assets/9b73a904ad91181a206d41ed4d4a5efa.png

Supplemental Table 1. CLI_combinations and the hyper-parameters of the classifiers constructed by different machine learning algorithm from these CLI_combinations.
URL: https://asset.jmir.pub/assets/f02df481dbfeef48e5de1b62fba5a9d8.xlsx