

1  
2 **An Interpretable Machine Learning Framework for Accurate Severe vs Non-severe**  
3 **COVID-19 Clinical Type Classification**

4  
5 **Authors:** Yuanfang Chen<sup>1,2†</sup>, Liu Ouyang<sup>3†</sup>, Forrest Sheng Bao<sup>4†</sup>, Qian Li<sup>5</sup>, Lei Han<sup>2,6</sup>, Baoli  
6 Zhu<sup>2,6,7</sup>, Ming Xu<sup>2,6,8\*</sup>, Jie Liu<sup>9</sup>, Yaorong Ge<sup>10</sup>, and Shi Chen<sup>8,11\*</sup>

7 **Affiliations:**

- 8 1. Public Health Research Institute of Jiangsu Province, Nanjing, 210009, China;  
9 2. Institute of HIV/AIDS/STI Prevention and Control, Jiangsu Provincial Center for Disease  
10 Control and Prevention, Nanjing, 210009, China;  
11 3. Department of Orthopedics, Union Hospital, Tongji Medical College, Huazhong University of  
12 Science and Technology, Wuhan, 430022, China;  
13 4. Department of Computer Science, Iowa State University, Ames, IA, 50011, USA;  
14 5. Department of Pediatrics, Affiliated Kunshan Hospital of Jiangsu University, Kunshan,  
15 215300, China;  
16 6. Department of Occupational Disease Prevention, Jiangsu Provincial Center for Disease  
17 Control and Prevention, Nanjing, 210009, China;  
18 7. School of Public health, Nanjing Medical University, Nanjing, 211166, China;  
19 8. Department of Public Health Sciences, College of Health and Human Services, University of  
20 North Carolina Charlotte, Charlotte, NC 28223, USA;  
21 9. Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of  
22 Science and Technology, Wuhan, 430022, China;  
23 10. Department of Software and Information Systems, College of Computing and Informatics,  
24 University of North Carolina Charlotte, Charlotte, NC 28223, USA;  
25 11. School of Data Science, University of North Carolina Charlotte, Charlotte, NC 28223, USA.

26  
27 \* Corresponding author: [mxu@uncc.edu](mailto:mxu@uncc.edu) (M.X.); [schen56@uncc.edu](mailto:schen56@uncc.edu) (S.C.)

28 † These authors contributed equally to this work.

29  
30 **One Sentence Summary:** We trained and validated machine learning random forest (RF)  
31 models to predict COVID-19 severity based on 26 comorbidity/symptom features and 26  
32 biochemistry features from a cohort of 214 non-severe and 148 severe type COVID-19 patients,  
33 identified top features from both feature modalities to differentiate clinical types, and achieved  
34 predictive accuracy of >90%, >95%, and >99% when comorbidity/symptom, biochemistry, and  
35 combined top features were used as input, respectively.

36

37  
38 **Abstract:** Effectively and efficiently diagnosing COVID-19 patients with accurate clinical type  
39 is essential to achieve optimal outcomes for the patients as well as reducing the risk of  
40 overloading the healthcare system. Currently, severe and non-severe COVID-19 types are  
41 differentiated by only a few clinical features, which do not comprehensively characterize  
42 complicated pathological, physiological, and immunological responses to SARS-CoV-2 invasion  
43 in different types. In this study, we recruited 214 confirmed COVID-19 patients in non-severe  
44 and 148 in severe type, from Wuhan, China. The patients' comorbidity and symptoms (26  
45 features), and blood biochemistry (26 features) upon admission were acquired as two input  
46 modalities. Exploratory analyses demonstrated that these features differed substantially between  
47 two clinical types. Machine learning random forest (RF) models using features in each modality  
48 were developed and validated to classify COVID-19 clinical types. Using comorbidity/symptom  
49 and biochemistry as input independently, RF models achieved >90% and >95% predictive  
50 accuracy, respectively. Input features' importance based on Gini impurity were further evaluated  
51 and top five features from each modality were identified (age, hypertension, cardiovascular  
52 disease, gender, diabetes; D-Dimer, hsTNI, neutrophil, IL-6, and LDH). Combining top 10  
53 multimodal features, RF model achieved >99% predictive accuracy. These findings shed light on  
54 how the human body reacts to SARS-CoV-2 invasion as a unity and provide insights on  
55 effectively evaluating COVID-19 patient's severity and developing treatment plans accordingly.  
56 We suggest that symptoms and comorbidities can be used as an initial screening tool for triaging,  
57 while biochemistry and features combined are applied when accuracy is the priority.  
58

## 59 Introduction

60 COVID-19 is a pandemic caused by the novel SARS-CoV-2 virus. As of May 17 2020, it  
61 has spread through at least 220 countries and regions, causing more than 4 million cases with  
62 300 thousand casualties (1). It is considered as the single most severe outbreak in the entire  
63 world during the 21st century, dwarfing other coronavirus-caused 2003 SARS and 2012 MERS  
64 epidemics. COVID-19 is especially challenging to the health professionals and general  
65 population. Unlike the precedent SARS and MERS epidemics, COVID-19 patients can be either  
66 asymptomatic or symptomatic, both of which are demonstrated to be transmissible of the virus  
67 with varying degrees (2-5). In addition, the distinct clinical types, non-severe and severe, require  
68 different treatment and care plans (6). Current studies are able to differentiate COVID-19  
69 patients from non-patients, but further detecting non-severe or severe types of COVID-19 is not  
70 comprehensively explored. Non-severe type patients can be accommodated in the mobile cabin  
71 hospital which requires relatively less intensive clinical monitoring and intervention, including  
72 treating pre-existing comorbidities, preventing healthcare associated infections and other  
73 comorbidities (8). In contrast, severe type patients need close monitoring, usually in ICU with  
74 more clinicians (6). Therefore, effectively and efficiently classifying COVID-19 clinical types is  
75 essential for triage, resource optimization, and care planning for front-line clinicians, healthcare  
76 systems, as well as for the patients (6,7).

77 Currently, non-severe and severe type are classified based on only a few clinical features  
78 (shortness of breath, O<sub>2</sub> saturation, and PaO<sub>2</sub>), which do not comprehensively characterize the  
79 complicated pathological, physiological, and immunological profile between non-severe and  
80 severe types in COVID-19 patients (9-11). In addition, some severe patients may not present  
81 shortness of breath initially. But without proper medical intervention, their clinical course will  
82 worsen abruptly, often resulting in respiratory failure with high mortality (6). It is therefore  
83 critical to provide accurate and efficient diagnosis of COVID-19 patients with correct clinical  
84 type information. We suggest that clinical features, including patient's comorbidities (e.g.,  
85 hypertension and diabetes), clinical symptoms (e.g., fever and chest pain), and blood  
86 biochemistry, are able to provide a more comprehensive characterization of COVID-19 and  
87 differentiate its clinical types (12,13). The human body is a unified and integrated entity. When  
88 pathogens such as SARS-CoV-2 invade, its effects can be shown not only from CT scans in the  
89 thoracic region, but also from other aspects such as clinical symptoms and biochemistry. ACE-2  
90 receptors, which facilitate SARS-CoV-2 infiltration, are distributed across multiple organs and  
91 systems in human body (35). More recent discoveries have found that in addition to respiratory  
92 system, SARS-CoV-2 can also invade digestive, reproductive, and even neural systems as well  
93 (14-17). In other words, comorbidities, clinical symptoms, and blood biochemistry information  
94 of COVID-19 patients could all be consequences and/or risk factors of SARS-CoV-2 infection.  
95 In clinical practice against COVID-19, clinicians not from respiratory or intensive care units may  
96 rely only on the referenced symptoms and signs (9) while neglecting diverse and important  
97 clinical features of COVID-19 patients, and may miss the critical signs of clinical course, leading  
98 to undesirable clinical consequences.

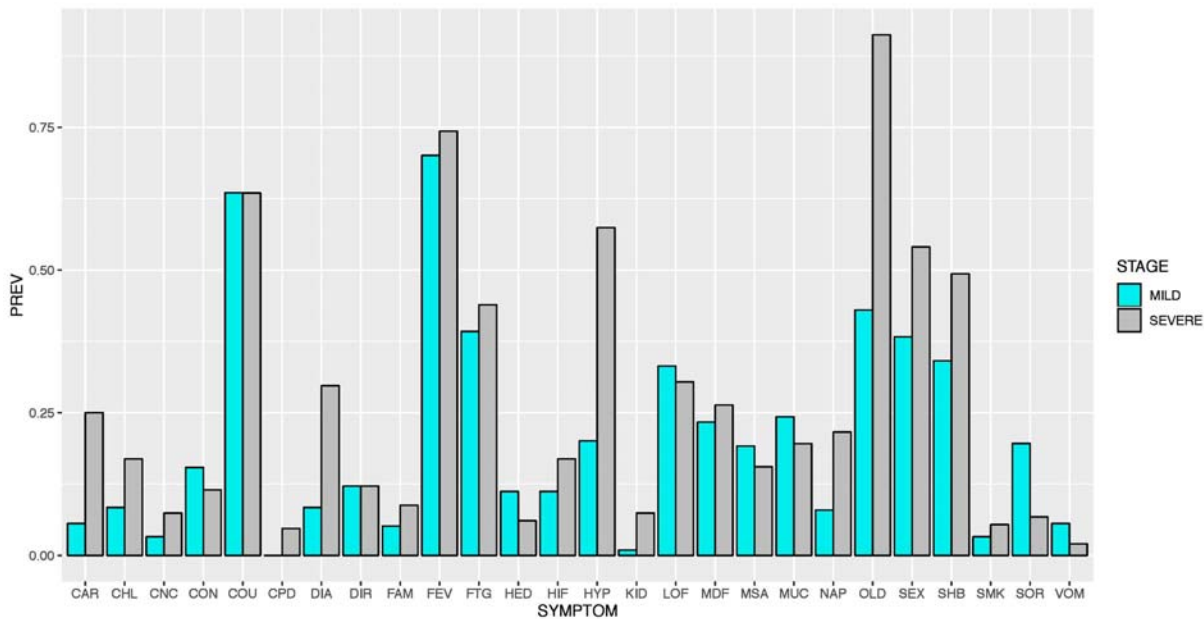
99 The potential power of symptoms, blood biochemistry, as well as their combinations to  
100 determine COVID-19 clinical type is currently not well understood nor evaluated (18-20). In  
101 order to utilize such diverse multimodality clinical information to make accurate and  
102 interpretable classifications, we propose a data mining and machine learning (ML) framework  
103 alternative to commonly used hypothesis-driven parametric models such as logistic regression.  
104 The results can provide reliable diagnostic decision support for clinicians even without

105 comprehensive experience on the emerging COVID-19. We aim to explore and contrast the  
106 distributions of comorbidities and symptoms, as well as blood biochemistry between non-severe  
107 and severe COVID-19 types. We will identify key features that differed substantially between  
108 the two clinical types and provide clear evidence-based interpretations for clinicians and other  
109 health professionals. Next, we will investigate whether single modality or specific combination  
110 of features across modalities are able to provide accurate classification models based on ML  
111 techniques. This study delivers an accurate diagnostic decision support tool to differentiate non-  
112 severe from severe type patients based on commonly available clinical data with clear clinical  
113 interpretations. Insights gained from this study, as well as developed end-to-end multimodal data  
114 analysis and ML framework, will enable us to better understand the comprehensive pathology of  
115 COVID-19, further distinguish COVID-19 from other infectious respiratory diseases, and apply  
116 in other diseases with multimodal clinical data in the future.

## 117 **Results**

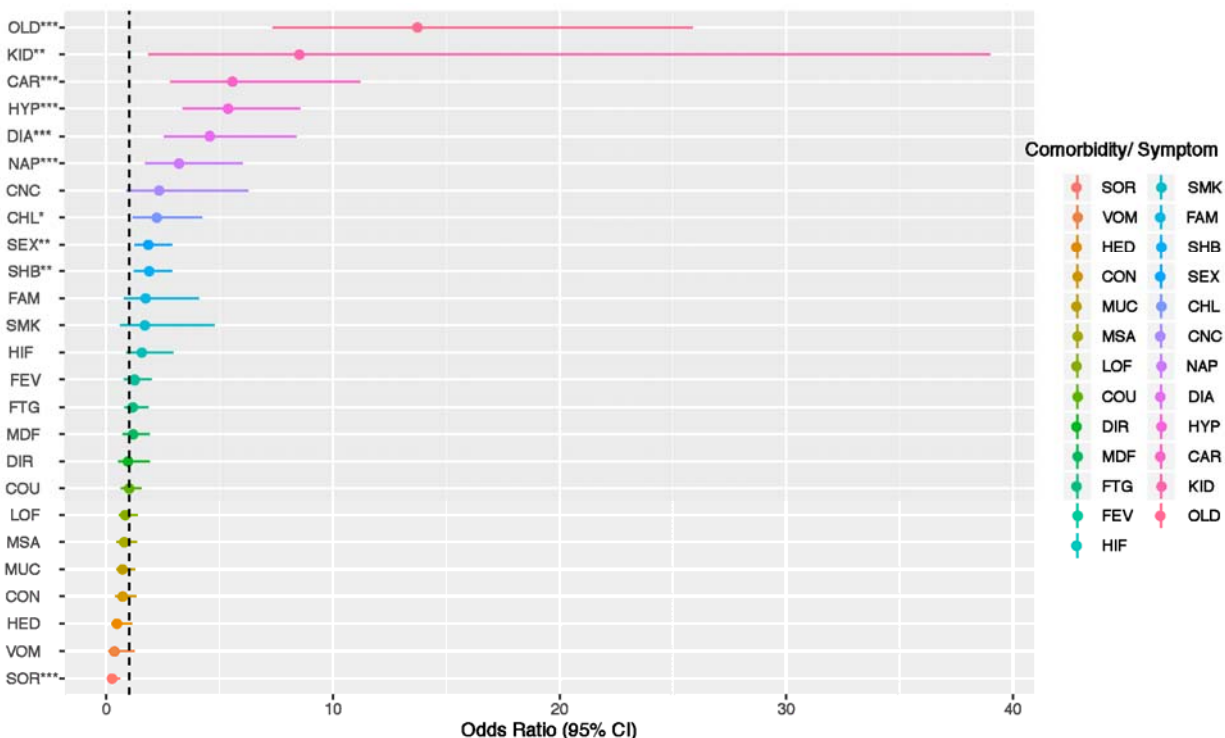
### 118 **Data Mining of COVID-19 Clinical Features between Non-severe and Severe Types**

119 Prevalence of symptom features in non-severe and severe COVID-19 types were  
120 calculated and compared (Fig. 1). Patients in the two clinical types showed distinct prevalence of  
121 many features. Severe COVID-19 patients were statistically much more likely to be elderly (at  
122 least age of 50, symbol OLD, OR=13.77, 95% CI= 7.33-25.86,  $p<0.001$ ) and male (SEX,  
123 OR=1.89, 95% CI= 1.24-2.90,  $p<0.01$ ), to have renal diseases (KID, OR=8.51, 95% CI= 1.86-  
124 38.99,  $p<0.001$ ), cardiovascular diseases (CAR, OR=5.61, 95% CI=2.81-11.20,  $p<0.001$ ),  
125 hypertension (HYP, OR=5.37, 95% CI=3.36-8.56,  $p<0.001$ ), diabetes (DIA, OR=4.61, 95%  
126 CI=2.53-8.38,  $p<0.001$ ), loss of appetite and taste (NAP, OR=3.20, 95% CI=1.70-6.01,  
127  $p<0.001$ ), feeling chilly (CHL, OR=2.21, 95% CI=1.16-4.22,  $p<0.05$ ), and chest congestion  
128 (SHB, OR=1.88, 95% CI=1.22-2.89,  $p<0.01$ ) than their non-severe counterparts. The only  
129 exception was sore throat, where severe patients had significantly much less likelihood to  
130 develop (SOR, OR=0.30, 95% CI=0.14-0.61,  $p<0.001$ ). Since some of these clinical conditions  
131 were self-reported symptoms, observation biases were to be expected. These discoveries were  
132 further demonstrated in the forest plot of odds ratio (OR) and CI in Fig. 2, showing the  
133 differences between the two clinical types. Therefore, these relatively easily measured and  
134 acquired clinical features could be utilized to clinically evaluate COVID-19 patients' severity.  
135 Our findings also echoed the U.S. CDC's recently updated list of symptoms of COVID-19 (21)  
136 and more recent reports on characterizations of COVID-19 patients in the U.S. (36). Our findings  
137 showed that elderly male COVID-19 patients with cardiovascular, respiratory, renal diseases and  
138 diabetes were at much higher risk of developing serious complications of COVID-19 such as  
139 acute respiratory distress syndrome (ARDS) and even death (19,20). In addition, we discovered  
140 that Chinese patients with renal diseases were significantly more likely to develop severe  
141 COVID-19, which was not widely reported before. Clinical evidence showed that ACE-2  
142 expression was associated with kidney diseases, thus making kidney disease a potential  
143 complication of SARS-CoV-2 invasion (22,23). This finding would inform clinicians to monitor  
144 kidney dysfunction, e.g., acute kidney injury, as a clinical sign and/or consequence of severe  
145 COVID-19 complication as well.



146  
147 **Fig. 1. Symptom Features Comparison between Non-severe and Severe Types**  
148 Note: symptom features were binary, so Y-axis was the prevalence of positives.

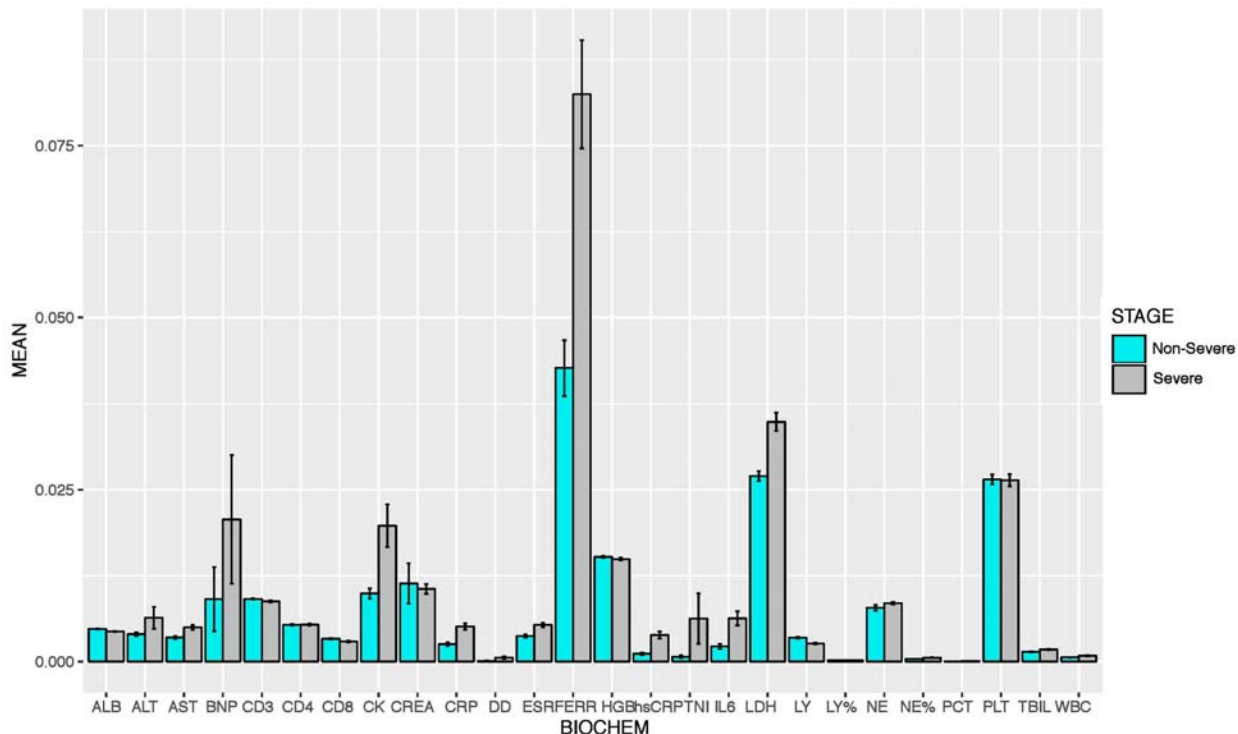
149  
150



151  
152 **Fig. 2. Forest Plot of Symptom Features between Non-severe and Severe Types**  
153 Note: \*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$  from the 2x2 contingency table for each feature. COPD  
154 was removed from the list because only severe type COVID-19 patients showed comorbidity of  
155 COPD. The threshold for a feature to be “positively” or “negatively” associated with severe

156 COVID-19 was 1 (dashed line), not 0. Forest plot is based on parametric statistical analysis and  
157 is irrelevant to random forest, a type of machine learning model used later in this study.  
158

159 For biochemistry modality features, we compared the actual distributions of these  
160 continuous features between non-severe and severe types. The results were demonstrated in Fig.  
161 4. Based on the two-sided Kolmogorov-Smirnov test results, severe and non-severe COVID-19  
162 types differed significantly in most biochemistry features, except platelet (PLT), hemoglobin  
163 (HGB), CD3, and CD4. Among all biochemistry features, IL-6, hsTNI, and D-dimer had the  
164 most significant differences between non-severe and severe COVID-19 types.  
165



166 **Fig. 3. Biochemistry Features Comparison between Non-severe and Severe Types**  
167 Note: values shown on y-axis were after feature scaling and were between 0 and 1. Error bars  
168 represented standard error (SE) of each biochemistry feature.  
169

170  
171 In addition, supplementary Fig. 1 showed symptom and biochemistry PCA results  
172 between non-severe and severe types. PCA plots reinforced the conclusion that associations  
173 among the features were substantially different between the two clinical types. The two types not  
174 only had vastly different distributions of features, interrelationships among features were also  
175 distinct between the two types.

176 In conclusion, after extensive clinical feature extraction and data mining, there were  
177 strong qualitative and quantitative evidences that non-severe and severe COVID-19 types  
178 differed substantially with regard to comorbidities, symptoms, and blood biochemistry. These  
179 findings paved the way toward an effective machine learning (ML) classifier to accurately  
180 differentiate these two types in clinical practice.

181 **Clinical Type Classification via Machine Learning (ML): Comorbidity and**  
182 **Symptom Modality**

183 We first explored whether the relatively simple binary symptom features could provide  
 184 accurate insights on COVID-19 severity. Model performance was summarized in Table 1 upper  
 185 section. Based on 100 independent runs, the RF model reached an average of >99% and 92%  
 186 accuracy for training and testing sets, respectively (Table 1). AUC was 90.2% (82.9%-97.6%)  
 187 based on the receiver operating characteristic (ROC) curve (Fig. 4 left panel). The model  
 188 performed better in detecting true positives (i.e., severe type) than true negatives (i.e., non-severe  
 189 type). In other words, symptom features alone in RF models almost never falsely predicted  
 190 severe case as non-severe case, but with a higher chance to predict non-severe case to severe  
 191 case. In clinical practice, this would be a lesser concern, as false positive (failed to detect mild  
 192 type) would be more tolerable than false negative (failed to detect severe type).

193 Our RF model also provided the major influential features to differentiate COVID-19  
 194 types based on contribution to Gini impurity. Top influential features were age, gender,  
 195 hypertension, diabetes, and cardiovascular diseases, in accordance with existing literature (24).  
 196 Other important symptom features included fatigue, chest congestion, sore throat, phlegm, and  
 197 fever. Most of these findings aligned well with our parametric data mining with odds ratio (OR)  
 198 comparison (Fig. 2, supplementary Table S1) but with much higher accuracy (90% accuracy on  
 199 prediction set of RF model compared to 68% accuracy from non-ML logistic regression). The  
 200 only exception was renal disease. While its prevalence was significantly different between the  
 201 two types, the RF model did not consider it as a major differentiating factor based on Gini  
 202 impurity (Supplementary Table S1). Clinically, elderly male patients with pre-existing  
 203 comorbidities, especially hypertension, diabetes, cardiovascular diseases were much more  
 204 vulnerable to COVID-19 and had a much higher risk to develop to severe type (18,20).  
 205 Therefore, we suggested using COVID-19 patients' comorbidity and symptom features as the  
 206 first round of evaluation of severity with reasonably high accuracy.

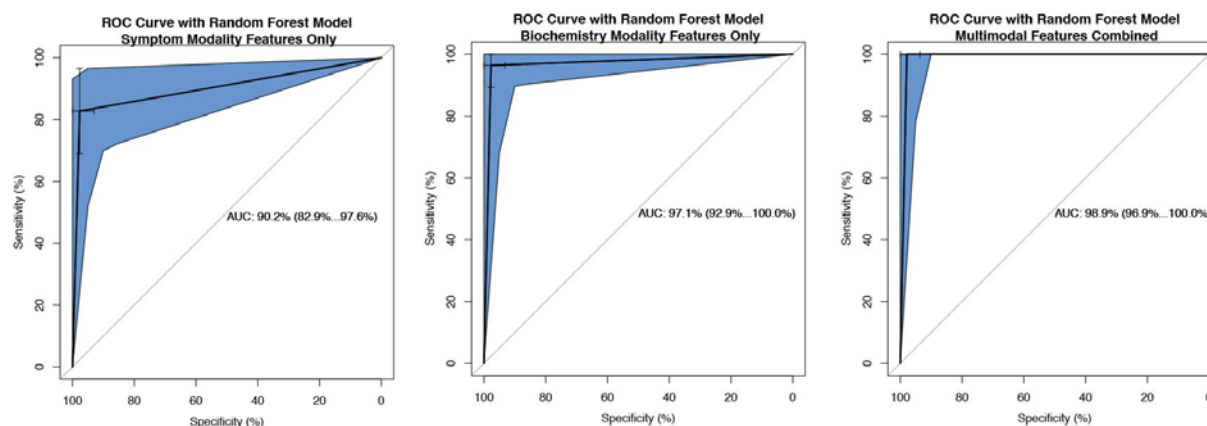
207  
 208

**Table 1. Random Forest Model Prediction Performance with Multimodal Features**

<b>Feature: Symptom</b>	<b>Median</b>	<b>Minimum</b>	<b>Maximum</b>
<b>Accuracy%</b>	90.28	83.33	98.61
<b>Sensitivity%</b>	97.5	87.80	>99
<b>Specificity%</b>	81.48	60.00	>99
<b>F1 Score%</b>	93.31	84.62	98.97
<b>AUC%</b>	90.20	82.90	97.60
<b>Feature: Biochemistry</b>	<b>Median</b>	<b>Minimum</b>	<b>Maximum</b>
<b>Accuracy%</b>	97.22	94.44	>99
<b>Sensitivity%</b>	97.92	91.11	>99
<b>Specificity%</b>	96.97	87.5	>99
<b>F1 Score%</b>	97.89	95.35	>99

<b>AUC%</b>	97.10	92.90	>99
<b>Features: Multimodal</b>	<b>Median</b>	<b>Minimum</b>	<b>Maximum</b>
<b>Accuracy%</b>	>99	97.22	>99
<b>Sensitivity%</b>	>99	97.22	>99
<b>Specificity%</b>	>99	92.00	>99
<b>F1 Score%</b>	>99	97.22	>99
<b>AUC%</b>	98.90	96.90	>99

209  
 210 Note: result based on 100 runs. Each run randomly selected 80% data as training set and 20% as  
 211 prediction set. Table shows model performance only on prediction set.  
 212  
 213



214  
 215 **Fig. 4. ROC Curve from Random Forest Model Based on Symptom, Biochemistry, and**  
 216 **Multimodal Features**

217 Note: panel (A) left showed symptom feature as input alone, panel (B) middle showed  
 218 biochemistry as input alone, and panel (C) right showed both features combined as input.

219  
 220 **Clinical Type Classification via ML: Biochemistry Modality**

221 Similar to symptom modality, RF model achieved an excellent performance in  
 222 differentiating non-severe and severe types using 26 features from biochemistry modality. On  
 223 average, the RF model achieved >99% and >95% accuracy for training and testing sets,  
 224 respectively. Sensitivity, specificity, and F1 scores were all above 95%, using only 8 trees in the  
 225 RF model (Table 1, middle section). AUC was 98% based on ROC curve (Fig. 4 middle panel).  
 226 Though this study focused on ML methods, we evaluated model performance of non-ML logistic  
 227 regression in supplementary Table S3 as a reference point to show the improvement that state-of-  
 228 the-art ML models could achieve.

229 Top differentiating features in biochemistry modality were D-dimer (DD), high  
 230 sensitivity troponin I (hsTNI), neutrophil (NE), interleukin-6 (IL-6), lactate dehydrogenase  
 231 (LDH), and high sensitivity c-reactive protein (hsCRP). The clinical interpretation of their



232 important role was that severe COVID-19 patients had more intensive immune response and  
233 hyperinflammation, such as cytokine storm syndrome with substantially increased IL-6 (25).  
234 Research also showed that SARS-CoV-2 was able to infect many organs other than lungs and  
235 induce dysfunction of these organs, including heart (26,27). Increasing hsTNI was a sign of heart  
236 tissue damage from SARS-CoV-2 infection (28). In addition, severe COVID-19 patients might  
237 have formed microthrombosis which induced higher D-dimer (18, 20,29-31). Abnormal level of  
238 neutrophils could be responsible for cytokine storm and ARDS in severe COVID-19 patients  
239 (13,32). hsCRP, a biomarker of acute inflammation, cardiovascular disease, and ischemic events,  
240 was also confirmed as the major contributing factor of COVID-19 mortality (18). LDH was a  
241 biomarker of tissue damage and was used to predict the clinical course of COVID-19 patients  
242 (42). These findings added further clinical insights in how multiple organs and systems, not just  
243 lungs, responded to SARS-CoV-2 infection in different clinical types (33-35).

244 Therefore, RF models developed in this study provided both high accuracy and valuable  
245 insights to identify clinical differences between COVID-19 types as well.

#### 246 **Clinical Type Classification via ML: Multimodal Features**

247 Based on the success of single modality RF models, we further developed a multimodal  
248 RF model that incorporated features and insights from both modalities. Instead of putting every  
249 feature in each modality, we only selected top 5 features from symptoms and top 5 features from  
250 biochemistry modalities. The results were encouraging and promising: these top 10 features out  
251 of a total of 52 features from both modalities achieved >99% in every model performance metric,  
252 including accuracy, sensitivity, specificity, and F1 score (Table 1, 3rd section). AUC was >99%  
253 as well (Fig. 4 right panel). Therefore, we suggested a two-step evaluation process of COVID-19  
254 patient's severity in clinical practice. Biochemistry and multimodal features such as a  
255 combination of symptom and biochemistry would serve as a more robust second-round  
256 confirmation after the first round of initial screening based on binary comorbidity and symptom  
257 features.

258 These findings reinforced our argument that SARS-CoV-2 attacked multiple organs and  
259 systems, and the human body reacted in a unity against its invasion. Different clinical features  
260 (e.g., comorbidity, symptom, and biochemistry) complemented each other to provide a more  
261 comprehensive characterization of human body as a united entity, not just respiratory system,  
262 reacted to SARS-CoV-2 invasion (35). In addition, the decent model performance promised the  
263 feasibility of multimodal clinical data mining in detecting and differentiating non-severe from  
264 severe COVID-19 patients. Our work would help effectively optimize healthcare operation  
265 during the pandemic and avoid overloading the healthcare system (7).

#### 266 **Discussion**

267 This study provides a breakthrough in combining the power of multiple clinical features  
268 from different modalities to differentiate COVID-19 clinical types via machine learning  
269 techniques. Practically, it enables delivering a more effective and efficient COVID-19 clinical  
270 type diagnostic decision support system. It helps develop optimal treatment plans for the  
271 individual patient, for example, sending to a mobile cabin hospital or admitting to a hospital with  
272 ICU (8). In addition, it will enable triaging and more effectively optimize the healthcare system  
273 resources and staffing. Doing so will substantially reduce the risk of overloading the healthcare  
274 system by admitting all COVID-19 patients into the hospital, decrease potential healthcare-  
275 associated infections, and improve clinical outcome for the patients, especially during this  
276 COVID-19 pandemic (7).

277 In addition to accurately detecting vulnerable COVID-19 patients who are likely to be in  
278 severe type, this study also provides clinical insights on why these patients may have been in  
279 severe type. Machine learning (ML) models work directly with data and therefore are generally  
280 not good at providing clear interpretations. In this study, we combine the power of both  
281 hypothesis-driven and data-driven ML models. The most contributing comorbidity, symptom,  
282 and biochemical features help predict and explain potential COVID-19 clinical courses and  
283 prognosis. Our research echoes recent studies that characterize and predict clinical course,  
284 critical illness and mortality of COVID-19 patients (13,18,20). In particular, another decision  
285 tree-based algorithm (XGBoost) showed promising performance in predicting mortality of  
286 CoVID-19 patients (18). RF was technically similar to XGBoost and our results were consistent  
287 to identify the key differentiating biochemistry features: LDH, hsCRP, and lymphocyte.

288 A continuous-valued risk score calculator for predicting risk of transitioning to critical  
289 type (an even more severe type which requires ICU, invasive ventilator, or ECMO, and has a  
290 mortality rate as high as 50%) has been developed for COVID-19 patients (20). As a  
291 comparison, although our RF model predicts a 0-1 binary outcome for non-severe and severe  
292 type patients, the internal RF modeling process through decision tree approach actually  
293 calculates an intermediate score between 0 and 1. By using a cut-off threshold, the RF model  
294 reports a final dichotomized 0-1 outcome. Therefore, our analytical framework can be readily  
295 adjusted to provide a continuous risk score for clinical evaluation and triaging of COVID-19  
296 patients as well, if needed.

297 Many severe COVID-19 patients present symptoms in lungs, especially ground glass  
298 opaque (GGO), which can be detected by biomedical imaging techniques such as CT. However,  
299 a major clinical challenge of COVID-19 lies in the asymptomatic patient problem, thus making it  
300 far worse than other coronavirus epidemics including SARS and MERS. These patients showed  
301 little if not none of classic symptoms related to viral pneumonia, presented no GGO, yet they are  
302 almost as capable of transmitting the virus as symptomatic patients (4-6). We suggest that the  
303 term “asymptomatic” may be due to lack of a comprehensive evaluation and understanding of  
304 this novel pathogen and hosts’ pathophysiology, and not truly “asymptomatic”. By more  
305 extensive data mining we show that non-severe COVID-19 patients have many symptoms  
306 differently distributed than severe patients. Our study provides an alternative route to detect non-  
307 severe COVID-19 patients and complement current biomedical imaging procedures.

308 The next step of this study is to further include biomedical imaging modality. A technical  
309 barrier is that CT scan is a high-dimensional feature set while symptom and biochemistry have  
310 relatively low dimensionality. Therefore, CT scan, at its original form of imaging, cannot be  
311 effectively combined with other modalities. We will evaluate the feasibility of using convolution  
312 neural network (CNN, another time of ML technique) first to reduce feature space in CT scan  
313 and extracting a fully connected layer in CNN as a representation of CT scan feature. A fully  
314 connected layer is a 1-dimensional vector and has the same dimensionality with the other two  
315 modalities. Therefore, in theory we would be able to further combine CT scans with other  
316 clinical features and investigate the association between these features with regard to COVID-19.

317 COVID-19 is a complex disease where the pathogen not only attacks the respiratory  
318 system but other organs and systems that have ACE-2 receptors as well (35,36). Our findings  
319 reveal the complicated pathological, physiological, and immunological responses to SARS-CoV-  
320 2 invasion and shed light in understanding the complex interactions between the virus and human  
321 body. Though the multimodal data mining and ML framework is developed with severe/non-  
322 severe COVID-19 data, we suggest that the end-to-end framework is applicable to many disease

323 systems where multimodal inputs are common, including demographic information, comorbidity,  
324 biochemistry, imaging, and -omics data. Having a more holistic viewpoint and approach will  
325 enable us to understand and respond to these emerging diseases, especially the unprecedented  
326 COVID-19, more readily in the field. We will further explore this analytical framework, and  
327 transfer insights for future clinical studies such as differentiating healthy, non-COVID viral  
328 pneumonia, non-severe, and severe COVID-19 patients.

329 In this study, we recruited participants from a single hospital in Wuhan, the first epicenter  
330 of COVID-19. There will inevitably be selection bias, as currently the ethnicity group is limited  
331 to Chinese participants. Therefore, we want to inform our colleagues across the world and see  
332 whether different demographic backgrounds influence feature distributions between non-severe  
333 and severe types in the patients. Our findings have already been independently identified in  
334 COVID-19 patients across ethnicity groups (21, 28). Another pitfall we should be aware of is  
335 that comorbidities may be the consequence of SARS-CoV-2 invasion, or risk factors that  
336 increase the risk of infection. Although the participants' comorbidity, symptom, and  
337 biochemistry were evaluated upon admission to hospital, they could have been exposed to the  
338 pathogen long before hospitalization, given the long "asymptomatic" type of COVID-19. For  
339 example, it is unclear whether kidney damage in a patient is a risk factor to induce severe type  
340 COVID-19, or SARS-CoV-2 attacks the kidney and causes kidney damage (35). The causal  
341 relationship needs to be more systematically evaluated with carefully designed prospective  
342 cohort studies. Nevertheless, in clinical practice, observing renal diseases in COVID-19 patients  
343 would trigger an alarm of clinical course to severe type, and inform clinicians to take actions to  
344 prevent acute kidney failure and even death.

345 Additionally, different subtypes of the virus, their specific pathogenicity and virulence,  
346 and host-pathogen interactions, should also be taken into consideration when conducting and  
347 comparing studies across different regions of the world. The other factors that this study did not  
348 include are behavioral and societal aspects, for instance, whether and how utilizing mobile cabin  
349 hospitals to treat non-severe type patients reduce the rate of transition to severe type. COVID-19,  
350 like all other infectious diseases, has individual clinical, epidemiological, behavioral as well as  
351 societal factors during its epidemic. Therefore, we will also explore cross-scale individual  
352 clinical course and population-level epidemics in future studies.

## 353 **Materials and Methods**

### 354 **Data Source and Clinical Feature Extraction**

355 In this study, we recruited 362 COVID-19 patients, including 214 non-severe and 148  
356 severe patients, from Wuhan Union Hospital affiliated to Tongji Medical School, Huazhong  
357 University of Science and Technology, China. Definitions of non-severe and severe cases were  
358 mainly adopted from the official COVID-19 Diagnosis and Treatment Plan from the National  
359 Health Commission of China and consulted guidelines from American Thoracic Society as well  
360 (9-11). Patients in severe type should present any one of the following features: 1) respiratory  
361 rate > 30 breaths per minute; 2) oxygen saturation < 93% at rest; or 3) arterial oxygen partial  
362 pressure (PaO<sub>2</sub>)/fraction of inspired oxygen (FiO<sub>2</sub>) < 300mmHg (40kPa). Each COVID-19  
363 patient was confirmed by two independent qRT-PCR tests before admitted to this study. All  
364 patients signed informed consent forms before participation. Symptoms were evaluated and  
365 blood samples were drawn upon admission. No pediatric patients younger than 18 years old were  
366 admitted in this study. This study was evaluated and approved by the IRB committee of Union  
367 Hospital, Wuhan, China (approval number: 2020-IEC-J-345).

368 Patients' de-identified clinical information include two major modalities of features. The  
369 first modality was a total of 26 pre-existing comorbidities and clinical symptoms, colloquially  
370 referred to as symptom features hereinafter. These features included gender, age (dichotomized  
371 as elder and young using 50yr as a cut-off point), hypertension, coughing, different types of  
372 fever, etc. A detailed description of these 26 features was provided in supplementary table S1.  
373 All symptom features were coded as 0-1 binary variables.

374 In addition, we also collected patients' blood samples and performed blood chemistry  
375 testing. After initial screening, we excluded some features with too many missing data such as  
376 calcitonin. Oxygen saturation and PaO<sub>2</sub>, the severe and non-severe type defining features,  
377 according to the Diagnosis and Treatment Plan (9), were also excluded. There were 26  
378 biochemistry features in this study, including IL-6, hemoglobin, and various lymphocytes. A  
379 detailed description and units of these biochemistry features were provided in supplementary  
380 table 2. All these biochemistry features were continuous features, different from the binary  
381 features in the symptom modality.

### 382 **Data Mining on Multimodal Clinical Features**

383 First, we conducted data mining on the multimodal COVID-19 data. The original dataset  
384 had approximately 5% missing data and we used predictive mean matching (PMM) to impute the  
385 original data. PMM was a commonly used computational method to handle missing data. To  
386 evaluate the effectiveness of PMM, we used a subset of the original dataset with no data missing,  
387 randomly dropped 5% data to simulate potential data loss, re-extrapolated the data with PMM,  
388 and evaluated the mean square root error (RMSE) between the original and imputed datasets.  
389 The RMSE was less than 0.05, indicating the extrapolation was feasible and reliable. The  
390 imputed data were then passed on to successive data mining and machine learning steps.

391 For the 0-1 binary features in symptom modality, we calculated the prevalence of each  
392 feature, i.e., number of positives over the number of patients in each type, non-severe and severe.  
393 Z-test was then applied to investigate whether there was a statistically significant difference of  
394 prevalence of any binary feature between the two types. In addition, a forest plot of odds ratio  
395 (OR) and its 95% confidence interval (CI) of symptom features between severe and non-severe  
396 COVID-19 types was developed.

397 For the continuous features in biochemistry modality, we characterized and contrasted the  
398 distribution of each feature in both types. Because most features were not normally distributed,  
399 we applied a two-sided Kolmogorovs-Smirnov test instead of Student's *t*-test for each feature and  
400 investigated whether these features distributed differently between the two types. Additionally,  
401 principal component analysis (PCA) was applied to help visualize the distinction in feature  
402 distributions and associations between the two clinical types.

### 403 **COVID-19 Clinical Type Classification via Machine Learning**

404 Traditional hypothesis-driven parametric models, such as logistic regression, relied  
405 heavily on human decisions of how features interact with each other (i.e., interaction terms in  
406 logistic regression model), which might not reflect the underlying medical reality. In addition,  
407 these models had strict prerequisites to perform correctly, for instance, normality of residuals,  
408 homoscedasticity, and independence of input features. Initial exploratory analyses showed that  
409 input features in both symptom and biochemistry modalities were non-normality and high  
410 collinearity among the features. Another technical challenge to logistic regression in this study  
411 was a mixture of binary symptom and continuous biochemistry variables from two modalities.

412 Bearing these problems, logistic regression would not be a preferable modeling approach  
413 to accurately classify and predict COVID-19 clinical types. Our exploratory analysis showed that

414 logistic regression could only achieve 68% and 77% predictive accuracy on an 80-20 training-  
415 prediction split, using symptom and biochemistry features, respectively (supplementary Table  
416 S3). Thus, hypothesis-driven models such as logistic regression were less feasible in clinical  
417 settings requiring high accuracy, sensitivity, and specificity to differentiate COVID-19 types.

418 On the other hand, state-of-the-art machine learning (ML) classification models worked  
419 directly with the data to avoid human bias. In addition, ML models did not have restrictions on  
420 how input data should be distributed or related. Therefore, ML classification would be a more  
421 appropriate modeling approach to predict COVID-19 clinical type with complicated data  
422 structure in this study. Therefore, we developed an end-to-end ML framework to accurately  
423 predict COVID-19 patient's clinical type based on symptom and/or biochemistry modality  
424 features. We built random forest (RF) classification models, as RF was able to provide excellent  
425 interpretability of input variable's relative importance to support clinical decision-making. RF  
426 was a widely used ML model based on decision theory and decision tree approach. The internal  
427 validation process through bagging made RF especially accurate and reliable. RF was also robust  
428 against data loss and data unbalancing, e.g., more non-severe type patients than severe patients in  
429 our study (37-41). There were other types of ML classification methods though, for example, k-  
430 nearest neighbor, artificial neural network, and naive Bayes. However, the major goal of this  
431 study was not to compare performances of different ML models, we focused on RF to deliver the  
432 most accurate classification possible.

433 We assigned severe cases as "positive" and non-severe as "negative" in the classification.  
434 The goal of ML classification through RF was to accurately predict the patient's COVID-19 type,  
435 either "positive" (severe) or "negative" (non-severe), based on features from different clinical  
436 modalities. In this part of study, we first use a single modality of features as input. The detailed  
437 RF modeling and validation process were provided in supplementary material and method. We  
438 trained the model 100 independent times, each time with a randomly selected set of 80% data for  
439 training and the remaining 20% for prediction. Hyperparameter number of trees (ntree) in the RF  
440 model was set at a very low value (ntree=8) to avoid potential overfitting issues (38). Important  
441 ML performance metrics, including accuracy, sensitivity, specificity, F1 score, and area under  
442 curve (AUC) value based on receiver operating characteristic (ROC) curve were computed for  
443 the prediction set. In addition, RF was able to evaluate input variables' importance to  
444 differentiating the two types based on their contributions to Gini impurity (40). We further  
445 quantified input features' relative importance, identified top contributing features, and explored  
446 their clinical relevance and interpretability to COVID-19. The most important features to  
447 differentiate COVID-19 clinical types were also cross-checked with our results from exploratory  
448 data mining, including prevalence of symptom features and distribution of biochemistry features.

#### 449 **COVID-19 Clinical Type Classification with Multimodal ML**

450 In addition, we explored whether and how combining features across modalities  
451 improved classification performance. In non-ML methods, it would be difficult to combine 0-1  
452 binary inputs with continuous inputs. However, this challenge was non-existent in ML models  
453 because ML models worked directly with data without *a priori* assumptions of the data structure.

454 We developed another RF model that incorporated features from both modalities. The  
455 modeling process was exactly the same as using single modality. Nevertheless, instead of putting  
456 every feature into the model, we selected top 5 features from each of the two modalities as new  
457 inputs. These top features were identified from the Gini impurity of single modality RF models  
458 (supplementary Table S1 and S2). We explored whether a few important features from different

459 modalities could perform sufficiently well to address the clinical challenge of differentiating  
460 non-severe and severe COVID-19 patients.

461 All statistical analyses and ML models were built in R 3.5.0. and Python 3.7 with  
462 additional supporting packages. The codes and fully de-identified data would be freely available  
463 on GitHub (<https://github.com/forrestbao/corona/tree/master/blood>).

464

## 465 **References and Notes:**

- 466 1. World Health Organization. Coronavirus Disease 2019 (COVID-19) Situation  
467 Report 118, 17 May 2020 (2020). [https://www.who.int/docs/default-](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200517-covid-19-sitrep-118.pdf?sfvrsn=21c0dafa_6)  
468 [source/coronaviruse/situation-reports/20200517-covid-19-sitrep-118.pdf?sfvrsn=21c0dafa\\_6](https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200517-covid-19-sitrep-118.pdf?sfvrsn=21c0dafa_6)
- 469 2. M. Gandhi, D. S. Yokoe, D. V. Havlir, Asymptomatic Transmission, the Achilles' Heel of  
470 Current Strategies to Control Covid-19. *N. Engl. J. Med.* 10.1056/NEJMe2009758 (2020).
- 471 3. R. Wölfel, W. M. Corman, W. Guggemos W, et al. Virological assessment of hospitalized  
472 patients with COVID-2019. *Nature* (2020). <https://doi.org/10.1038/s41586-020-2196-x>
- 473 4. Y. Luo, E. Trevathan, Z. Qian, Y. Li, J. Li, W. Xiao, N. Tu, Z. Zeng, P. Mo, Y. Xiong, G.  
474 Ye, Asymptomatic SARS-CoV-2 infection in household contacts of a healthcare provider,  
475 Wuhan, China. *Emerg. Infect. Dis.* **26**, (2020). <https://doi.org/10.3201/eid2608.200282>
- 476 5. F. Ye, S. Xu, Z. Rong, R. Xu, X. Liu, P. Deng, et al., Delivery of infection from  
477 asymptomatic carriers of COVID-19 in a familial cluster. *Int. J. Infect. Dis.* **94**, 134-138  
478 (2020).
- 479 6. Z. Wu, J.M. Googan JM, Characteristics of and important lessons from the coronavirus  
480 disease 2019 (COVID-19) outbreak in China: summary of a report of 72,314 cases from the  
481 Chinese Center for Disease Control and Prevention. *JAMA* **323**, 1239 (2020).
- 482 7. R. D. Truog, C. Mitchell, G. Q. Dalley, The toughest triage—allocating ventilators in a  
483 pandemic. *N. Engl. J. Med.* (2020). <https://doi.org/10.1056/>
- 484 8. S. Chen, Z. Zhang, J. Yang, J. Wang, X. Zhai, T. Bornighausen, C. Wang, Fangcang  
485 shelterhospitals: a novel concept for responding to public health emergencies. *Lancet* **395**,  
486 1309-1313 (2020).
- 487 9. National Health Commission of China. Novel coronavirus pneumonia diagnosis and  
488 treatment plan (provisional 7<sup>th</sup> edition). 4 Mar 2020 (2020).  
489 <http://www.nhc.gov.cn/zwygj/s7652m/202003/a31191442e29474b98bfed5579d5af95.shtml>
- 490 10. J.P. Metlay, G.W. Waterer, A.C. Long, A. Anzueto, J. Brozek, K. Crothers, L. A. Cooley, N.  
491 C. Dean, M. J. Fine, S. A. Flanders, M. R. Griffin, Diagnosis and treatment of adults with  
492 community-acquired pneumonia. an official clinical practice guideline of the American  
493 Thoracic Society and Infectious Diseases Society of America. *Am. J. Respir. Crit. Care Med.*  
494 **200**, e45-e67 (2019).
- 495 11. A.M. Neill, I. R. Martin, R. Weir, R. Anderson, A. Cheresky, M. J. Epton, R. Jackson, M.  
496 Schousboe, C. Frampton, S. Hutton, S.T. Chambers, G.I. Town, Community acquired  
497 pneumonia: aetiology and usefulness of severity criteria on admission. *Thorax* **51**, 1010-1016  
498 (1996). doi:10.1136/thx.51.10.1010
- 499 12. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team, The  
500 epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-  
501 19) in China. *Zhonghua Liu Xing Bing Xue Za Zhi* **41**, 145–151 (2020).
- 502 13. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, et al., Clinical features of patients  
503 infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).

- 504 14. X. Liu. et al., A general report on the systematic anatomy of COVID-19. *J. Forensic Med.*  
505 **36**, 1–3 (2020).
- 506 15. G. Cholankeril, A. Podboy, V. I. Aivaliotis, B. Tarlow, E. A. Pham, S. Spencer, D. Kim, A.  
507 Hsing, A. Ahmed, High Prevalence of Concurrent Gastrointestinal Manifestations in Patients  
508 with SARS-CoV-2: Early Experience from California. *Gastroenterology* (2020).  
509 doi.org/10.1053/j.gastro.2020.04.008.
- 510 16. Y. Jing, L. Run-Qian, W. Hao-Ran, C. Haoran, L. Ya-Bin, G. Yang, C. Fei, Potential  
511 influence of COVID-19/ACE2 on the female reproductive system. *Mol. Hum. Reprod.*  
512 (2020). doi:10.1093/molehr/gaaa030.
- 513 17. T. Moriguchi, N. Harii, J. Goto, D. Harada, H. Sugawara, J. Takamino, et al., A first case of  
514 meningitis/encephalitis associated with SARS-Coronavirus-2. *Int. J. Infect. Dis.* **94**, 55-58  
515 (2020).
- 516 18. L. Yan, H. Zhang, J. Goncalves, et al., An interpretable mortality prediction model for  
517 COVID-19 patients. *Nat Mach Intell* (2020). <https://doi.org/10.1038/s42256-020-0180-7>
- 518 19. J. Liu, Longitudinal characteristics of lymphocyte responses and cytokine profiles in the  
519 peripheral blood of SARS-CoV-2 infected patients. *EbioMedicine* **55**, 102763 (2020).
- 520 20. W. Liang, H. Liang, L. Ou, B. Chen, A. Chen, C. Li, Y. Li, W. Guan, L. Sang, J. Lu, Y. Xu,  
521 G. Chen, H. Guo, J. Guo, Z. Chen, Y. Zhao, S. Li, N. Zhang, N. Zhong, J. He, Development  
522 and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in  
523 Hospitalized Patients With COVID-19. *JAMA Intern. Med.*, e202033 (2020).  
524 doi:10.1001/jamainternmed.2020.2033
- 525 21. Centers for Disease Control and Prevention. Symptoms of Coronavirus (COVID-19).  
526 <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>
- 527 22. S. Mizuiri, Y. Ohashi, ACE and ACE2 in kidney disease. *World J. Nephrol.* **4**, 74-82 (2015).
- 528 23. L. Lieben, ACE2 administration slows kidney damage. *Nat. Rev. Nephrol.* **13**, 261 (2017).
- 529 24. X. Yang, Y. Yu, J. Xu, H. Shu, J. Xia, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, Y.  
530 Wang, S. Pan, X. Zou, S. Yuan, Y. Shang, Clinical course and outcomes of critically ill  
531 patients with SARS CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective,  
532 observational study. *Lancet Resp. Med.* **8**, 475–481 (2020).
- 533 25. Y. Li, Y. Hu, J. Yu, et al., Retrospective analysis of laboratory testing in 54 patients with  
534 severe- or critical-type 2019 novel coronavirus pneumonia. *Lab Invest.* (2020).  
535 <https://doi.org/10.1038/s41374-020-0431-6>
- 536 26. L. Chen, X. Li, M. Chen, Y. Feng, C. Xiong, The ACE2 expression in human heart indicates  
537 new potential mechanism of heart injury among patients infected with SARS-CoV-2.  
538 *Cardiovascular Research*, **116**, 1097–1100 (2020).
- 539 27. S. Shi, M. Qin, B. Shen, et al., Association of Cardiac Injury With Mortality in Hospitalized  
540 Patients With COVID-19 in Wuhan, China. *JAMA Cardiol.* (2020).  
541 doi:10.1001/jamacardio.2020.0950
- 542 28. J.W. Li, T.W. Han, M. Woodward, et al., The impact of 2019 novel coronavirus on heart  
543 injury: A Systematic review and Meta-analysis. *Prog Cardiovasc Dis.* (2020)  
544 doi:10.1016/j.pcad.2020.04.008
- 545 29. C. Wu, X. Chen, Y. Cai, et al., Risk factors associated with acute respiratory distress  
546 syndrome and death in patients with coronavirus disease 2019 pneumonia inWuhan, China.  
547 *JAMA Intern. Med.* **13**, (2020) doi:10.1001/jamainternmed.2020. 0994

- 548 30. W.J. Guan, Z.Y. Ni, Y. Hu, et al., China Medical Treatment Expert Group for Covid-19.  
549 Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med.* **28**, (2020)  
550 doi:10.1056/NEJMoa2002032
- 551 31. D.O. Griffin, A. Jensen, M. Kha, J. Chin, K. Chin, J. Saad, R. Parnell, C. Awwad, D. Patel,  
552 Pulmonary Embolism and Increased Levels of d-Dimer in Patients with Coronavirus Disease.  
553 *Emerg. Infect. Dis.* **26**, (2020).
- 554 32. B. J. Barnes, J. M. Adrover, A. Baxter-Stoltzfus, et al., Targeting potential drivers of COVID-  
555 19: Neutrophil extracellular traps. *J Exp Med.* **6**. (2020) doi:10.1084/jem.20200652
- 556 33. Zaim S, Chong JH, Sankaranarayanan V, Harky A. COVID-19 and Multi-Organ Response.  
557 *Curr Probl Cardiol.* 100618 (2020). doi:10.1016/j.cpcardiol.2020.100618
- 558 34. T. Wang, Z. Du, F. Zhu, Z. Cao, Y. An, Y. Gao, et al., Comorbidities and multi-organ  
559 injuries in the treatment of COVID-19. *Lancet*, **395**, E52 (2020). doi:10.10106/S0140-  
560 6736(20)30558-4
- 561 35. M. Wadman, J. Couzin-Frankel, J. Kaiser, C. Maticic, How does coronavirus kill? Clinicians  
562 trace a ferocious rampage through the body, from brain to toes. *Science*, (2020).  
563 doi:10.1126/science.abc3208
- 564 36. S. Garg, L. Kim, M. Whitaker, et al., Hospitalization Rates and Characteristics of Patients  
565 Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14  
566 States, March 1–30, 2020. *Morb. Mortal. Wkly Rep.* **69**, 458-464 (2020).  
567 DOI: <http://dx.doi.org/10.15585/mmwr.mm6915e3>
- 568 37. J.R. Quinlan, Introduction to decision tree. *Machine Learning* **1**, 81–106 (1986).
- 569 38. T.K. Ho, Random decision forests. *Proceedings of the 3rd International Conference on*  
570 *Document Analysis and Recognition*, 278–282 (1995).
- 571 39. S. Lundberg, et al., From local explanations to global understanding with explainable AI for  
572 trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
- 573 40. L. Tolosi, T. Lengauer, Classification with correlated features: unreliability of feature  
574 ranking and solutions. *Bioinformatics* **27**, 1986–1994 (2011).
- 575 41. C. Strobl, A. Boulesteix, T. Augustin, Unbiased split selection for classification trees based  
576 on the Gini index. *Comput. Stat. Data An.* **52**, 483–501 (2007).
- 577 42. F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, et al., Clinical course and risk factors for  
578 mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.  
579 *Lancet*, **395**, 1054-1062 (2020).

#### 580 **Acknowledgments:**

581 This work is a tribute to the frontline clinicians and supporting staff who devoted their lives to  
582 combating this COVID-19 pandemic.

583 **Funding:** this study was jointly supported by the National Science Foundation for Young  
584 Scientists of China (81703201), the Natural Science Foundation for Young Scientists of Jiangsu  
585 Province (BK20171076), the Jiangsu Provincial Medical Innovation Team (CXTDA2017029),  
586 the Jiangsu Provincial Medical Youth Talent program (QNRC2016548), the Jiangsu Preventive  
587 Medicine Association program (Y2018086), the Lifting Program of Jiangsu Provincial Scientific  
588 and Technological Association, and the Jiangsu Government Scholarship for Overseas Studies.

589 **Author contributions:** Y.C. designed the study. L.O. and J.L. derived and processed data. L.O.,  
590 F.S.B., Q.L., L.H., B.Z, J.L., and S.C. interpreted results. M.X. and S.C. performed analyses.  
591 Y.G. and S.C. supervised. S.C. developed the manuscript. Y.C., L.O., and F.S.B. contributed  
592 equally.

593 **Competing interests:** the authors declare no competing interests in this study.



594 **Data and materials availability:** de-identified clinical data and codes in this study sare openly  
595 available on GitHub (<https://github.com/forrestbao/corona/tree/master/blood>)  
596

597

598 **Figure and Table Legends:**

599 **Fig. 1. Symptom Features Comparison between Non-severe and Severe Stage COVID-19**  
600 **Patients**

601 Note: symptom features were binary so Y-axis was the prevalence of positives.

602 **Fig. 2. Forest Plot of Top Symptom Features that Differ Significantly between Stages**

603 Note: \*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$  from the 2x2 contingency table. Forest plot is based on  
604 parametric statistical analysis and is irrelevant to random forest, a type of machine learning  
605 model used later in this study. The threshold for a feature to be “positively” or “negatively”  
606 associated with severe COVID-19 was 1 (dashed line), not 0.

607 **Fig. 3. Biochemistry Features Comparison between Non-severe and Severe Stage COVID-**  
608 **19 Patients**

609 Note: values shown on y-axis were after feature scaling and were between 0 and 1. Error bars  
610 represented standard error (SE) of each biochemistry feature.

611 **Fig. 4. ROC Curve from Random Forest Model Based on Symptom, Biochemistry, and**  
612 **Multimodal Features**

613 Note: panel (A) left showed symptom feature as input alone, panel (B) middle showed  
614 biochemistry as input alone, and panel (C) right showed both features combined as input.

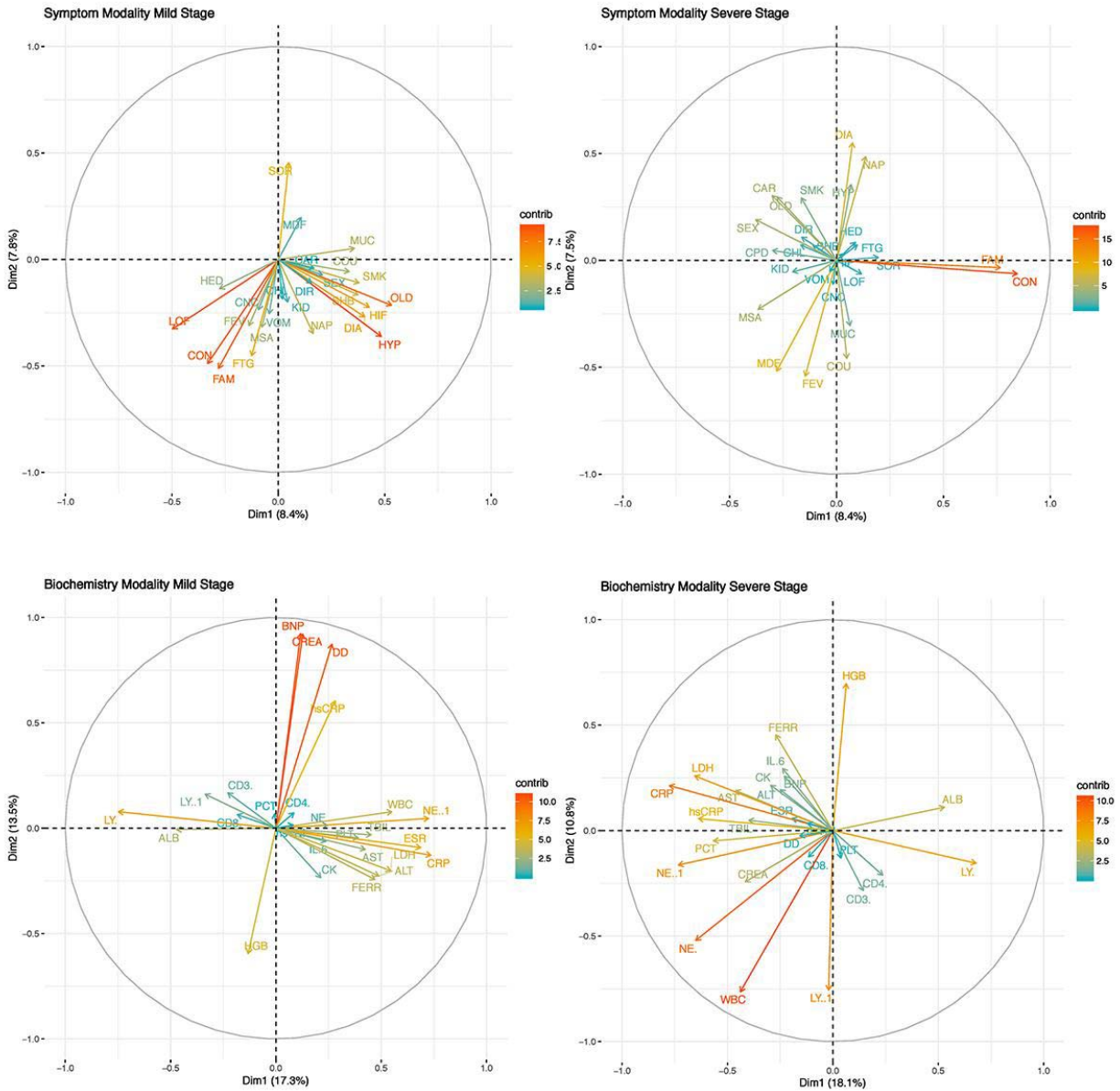
615 **Table 1. Random Forest Model Prediction Performance with Multimodal Features**

616

617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643

### **Supplementary Method: Random Forest Machine Learning Model Formulation**

In each run, we randomly selected 80% of all data as the training set to train the RF model. These 80% data included both severe and non-severe cases, i.e., both positives and negatives. We also ensured that the distributions of positives and negatives in the training set was similar to those in the complete data. Once the model was developed, the remaining 20% data would be fed into the developed model to evaluate its performance on unseen prediction data. This prediction process was crucial to ensure that the ML model was not over-fitting, i.e., the model worked extremely well on existing training data but poorly on unseen real-world data. We then constructed the 2x2 confusion matrix to evaluate the model performance on prediction data. The 2x2 confusion matrix had four elements, true positive (TP, model correctly identified severe type), true negative (TN, model correctly identified non-severe type), false positive (FP, model incorrectly identified non-severe type as severe type), and false negative (FN, model incorrectly identified severe type as non-severe). Then, important ML model performance metrics were computed, including model accuracy, sensitivity, specificity, and F1 score, etc. Among these performance metrics, accuracy and F1 score evaluated overall performance of the model, sensitivity (also known as true negative rate, TNR) emphasized FN, and specificity (also known as true positive rate, TPR) emphasized FP. Our RF model aimed to increase TP and TN while simultaneously reducing FP and FN. In the other words, an ideal ML model should have both high sensitivity and high specificity. The highest possible value for these metrics was 1 (100%), indicating the model could correctly distinguish all positive (severe) types from negative (non-severe) types. In this study, we run this modeling and predicting process 100 times to evaluate how system stochasticity influenced the RF model and whether the RF model performance was robust. In each of the 100 runs, a different set of randomly selected 80% data were used to train the model and the remaining 20% to predict and evaluate the model performance. We reported maximum, minimum, and median values of performance metrics (accuracy, F1 score, AUC, etc.).



644  
645 **Supplementary Fig. 1. PCA Plot for Features in Symptom and Biochemistry Modalities**  
646 **between Clinical Types**

647 Note: upper section: symptom modality; lower section: biochemistry modality; left section: non-  
648 severe (mild) COVID-19 type; right section: severe COVID-19 type.  
649

650  
651

**Table S1. Comorbidity and Symptom Features**

<b>Abbrev.</b>	<b>Health Condition/Symptom</b>	<b>Random Forest Gini Impurity (%)</b>	<b>Logistic Regression Coefficient</b>	<b>Note</b>
<b>OLD</b>	Elderly	24.94	1.85	Age>50 as elderly (OLD=1)
<b>HYP</b>	Hypertension	14.43	0.63	Diastolic>90 or systolic>140
<b>CAR</b>	Cardiovascular diseases	8.57	0.75	
<b>SEX</b>	Biological gender	7.79	0.63	Male=0, female=1
<b>DIA</b>	Diabetes	6.73	0.39	Type 2 diabetes only
<b>FTG</b>	Fatigue	6.33	0.32	Subjective, self-reported
<b>SHB</b>	Chest congestion	6.29	0.28	
<b>SOR</b>	Sore throat	5.92	-0.9	
<b>MUC</b>	Phlegm	5.63	-0.58	
<b>FEV</b>	Fever (any)	5.45	-0.91	>37C (>98.6F)
<b>COU</b>	Coughing	5.41	-0.05	
<b>MSA</b>	Muscle ache	5.39	-0.58	
<b>NAP</b>	Loss of appetite	5.22	0.77	
<b>CON</b>	Contacting COVID-19 patients	4.33	-0.28	
<b>MDF</b>	Medium fever	4.31	1.44	38.1-39C (100.5-102.2F)
<b>LOF</b>	Low fever	4.29	1.3	37.1-38C (98.7-100.4F)
<b>CHL</b>	Chilling and shaking	4.22	0.91	
<b>DIR</b>	Diarrhea	3.86	-0.45	
<b>HIF</b>	High fever	3.85	1.57	>39C (>102.2F)
<b>VOM</b>	Vomiting	2.79	-1.7	
<b>KID</b>	Renal diseases	1.84	1.73	
<b>HED</b>	Headache	1.8	0.07	Any type of headache
<b>CNC</b>	Cancer	1.78	0.16	Any type of cancer
<b>FAM</b>	Family members with COVID-19	1.54	1.11	
<b>SMK</b>	Smoking history	1.15	-0.75	
<b>CPD</b>	Chronic obstructive pulmonary disease (COPD)	1.08	16.28	

652  
653 **Note: bold** are top ten symptom features critical to differentiate COVID-19 non-severe and  
654 severe types from machine learning random forest model. Logistic regression coefficient signs  
655 (positive or negative) reveal if the feature increases or decreases risk of developing severe type  
656 COVID-19.

657 **Table S2. Blood Biochemistry Features**

Abbrev.	Blood Biochemistry	Importance by Gini Impurity (%)	Logistic Regression Coefficient	Unit and Note
<b>DD</b>	D-dimer	25.41	0.5	mg/L
<b>hsTNI</b>	High sensitivity Troponin I	16.06	0.0031	ng/mL
<b>LDH</b>	Lactate dehydrogenase	10.19	0.0012	U/L
<b>NE</b>	Neutrophil	10.02	0.0044	10 <sup>9</sup> /L
<b>IL6</b>	Interleukin 6	9.41	0.026	ng/mL
<b>hsCRP</b>	High sensitivity C-reactive protein	9.11	0.019	ug/L
<b>ESR</b>	Erythrocyte sedimentation rate	7.96	0.029	mm/h
<b>TBIL</b>	Total bilirubin	7.43	0.018	umol/L
<b>CD8</b>	Cluster of differentiation 8	7.09	-0.097	/uL
<b>CK</b>	Creatine kinase	6.7	<0.001	U/L
CRP	C-reactive protein	6.69	-0.02	ug/L
FER	Ferritin	5.86	<0.001	ug/L
ALT	Alanine transaminase	5.52	-0.0037	U/L
CREA	Creatinine	4.45	0.01	umol/L
LY%	Percent of Lymphocyte	4.4	0.91	%
CD3	Cluster of differentiation 3	4.06	0.051	/uL
ALB	Albumin	4.02	0.015	g/L
NE%	Percent of Neutrophil	3.94	0.51	%
PLT	Platelet	3.91	<0.001	10 <sup>9</sup> /L
AST	Aspartate aminotransferase	3.85	0.017	U/L
PCT	Procalcitonin	3.57	-0.28	ng/mL
CD4	Cluster of differentiation 4	3.51	-0.056	/uL
LY	Lymphocyte	3.26	-0.055	10 <sup>9</sup> /L
WBC	White blood cell	3.13	-0.4	10 <sup>9</sup> /L
BNP	Brain natriuretic peptide	3.07	0.0033	pg/mL
HGB	Hemoglobin	2.98	-0.024	g/L

658  
 659 **Note:** bold are top ten biochemistry features critical to differentiate COVID-19 non-severe and  
 660 severe types from machine learning random forest (RF) model.  
 661

662 **Table S3. Logistic Regression Prediction Performance**

<b>Feature</b>	<b>Symptom</b>	<b>Biochemistry</b>
<b>Accuracy%</b>	69.44	78.62
<b>Sensitivity%</b>	65.22	78.65
<b>Specificity%</b>	76.92	78.57
<b>F1 Score%</b>	71.07	78.61

663  
664 **Note:** multimodal logistic regression is not performed because two feature modalities are not  
665 compatible to combine, i.e., 0-1 binary symptom and continuous biochemistry input data.

666  
667  
668