EBioMedicine

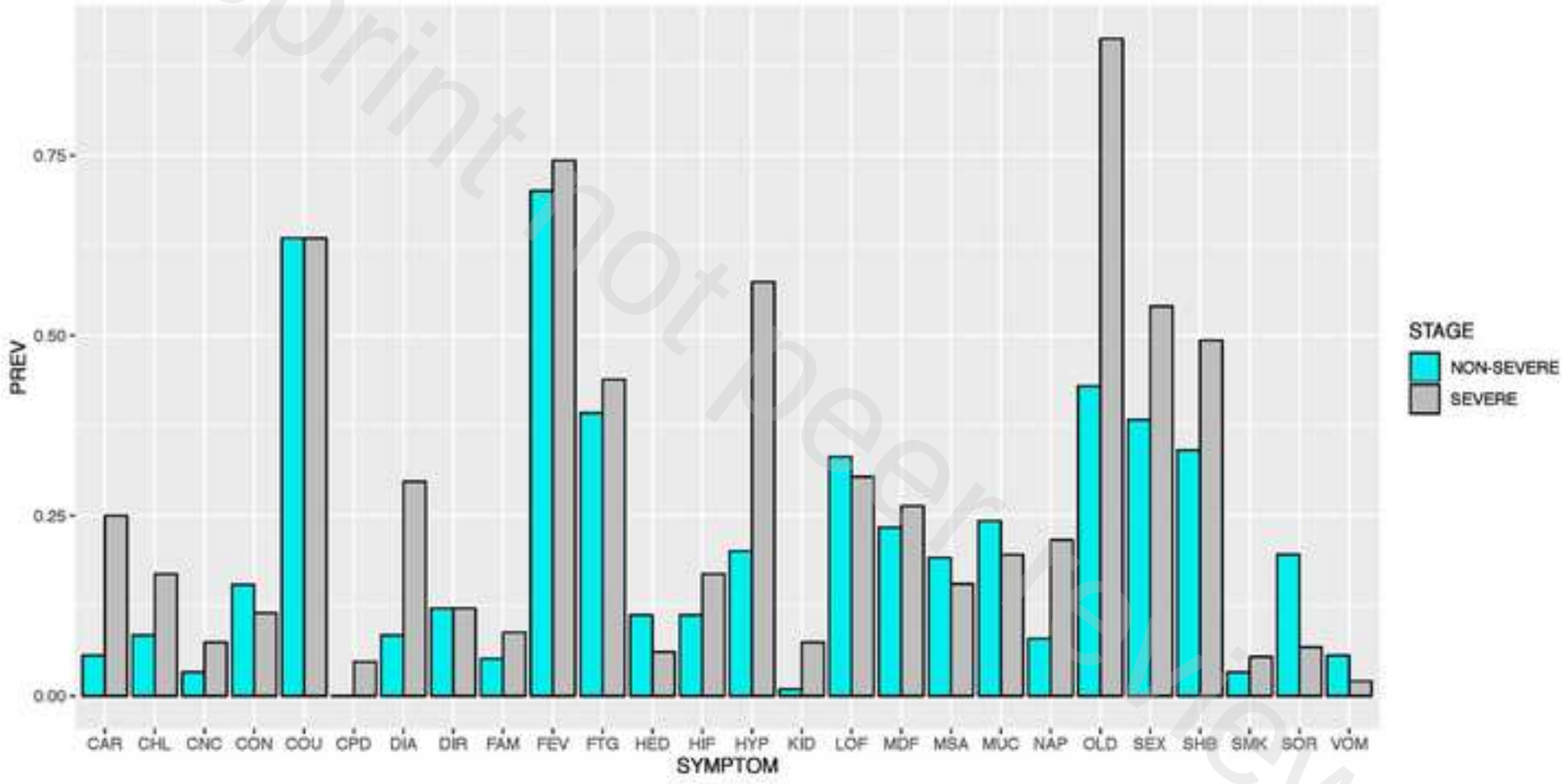
An Interpretable Machine Learning Framework for Accurate Severe vs Non-severe COVID-19 Clinical Type Classification

--Manuscript Draft--

Manuscript Number:	EBIOM-D-20-02109
Article Type:	Article (Original Research)
Keywords:	COVID-19, machine learning, supervised learning, clinical classification
Corresponding Author:	Ming Xu, Ph.D. Jiangsu Province Center for Disease Control and Prevention Nanjing, Jiangsu CHINA
First Author:	Yuanfang Chen
Order of Authors:	Yuanfang Chen
	Liu Ouyang
	Forrest Sheng Bao
	Qian Li
	Lei Han
	Hengdong Zhang
	Baoli Zhu
	Yaorong Ge
	Ming Xu, Ph.D.
	Jie Liu
	Shi Chen
Manuscript Region of Origin:	CHINA
Abstract:	Abstract : Effectively and efficiently diagnosing COVID-19 patients with accurate clinical type is essential to achieve optimal outcomes for the patients as well as reducing the risk of overloading the healthcare system. Currently, severe and non-severe COVID-19 types are differentiated by only a few features, which do not comprehensively characterize the complicated pathological, physiological, and immunological responses to SARS-CoV-2 invasion in different types. In this study, we recruited 214 confirmed COVID-19 patients in non-severe and 148 in severe type, from Wuhan, China. The patients' comorbidity and symptoms (including 26 features), and laboratory testing results (26 features) upon admission were acquired as two input modalities. Exploratory analyses demonstrated that these features differed substantially between two clinical types. Machine learning random forest (RF) models based on features in each modality were developed and validated to classify COVID-19 clinical types. Using comorbidity/symptom and laboratory results as input independently, RF models achieved >90% and >95% predictive accuracy, respectively. Input features' importance based on Gini impurity were further evaluated and top five features from each modality were identified (age, hypertension, cardiovascular disease, gender, diabetes; D-Dimer, hsTNI, absolute neutrophil count, IL-6, and LDH, in descending order). Combining top 10 multimodal features, RF model achieved >99% predictive accuracy. These findings shed light on how the human body reacts to SARS-CoV-2 invasion as a unity and provide insights on effectively evaluating COVID-19 patient's severity and developing personalized treatment plans accordingly. We suggest that symptoms and comorbidities can be used as an initial screening tool for triaging, while laboratory results are applied when accuracy is the priority.

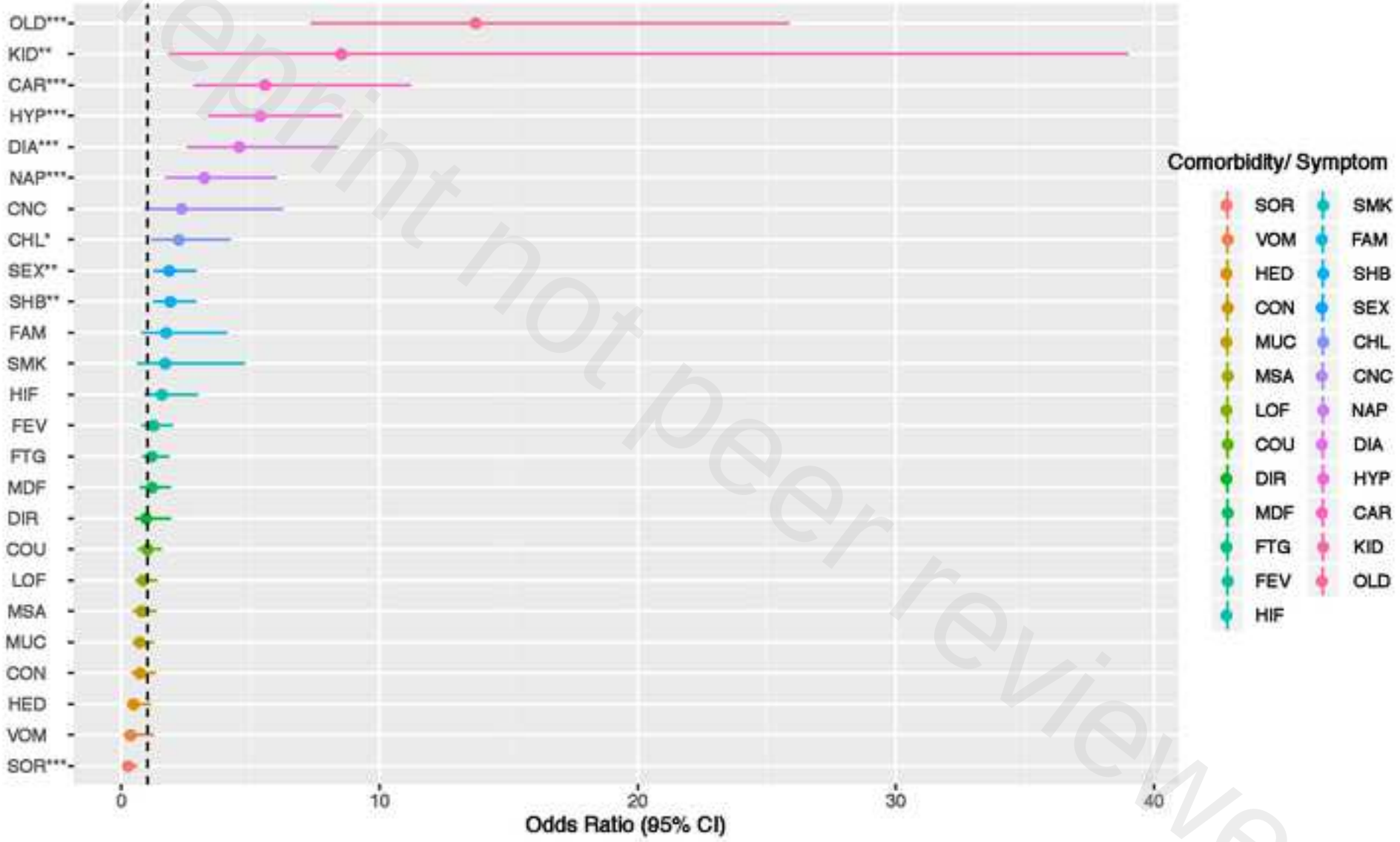
Figure

Click here to access/download;Figure;figure 1.tif



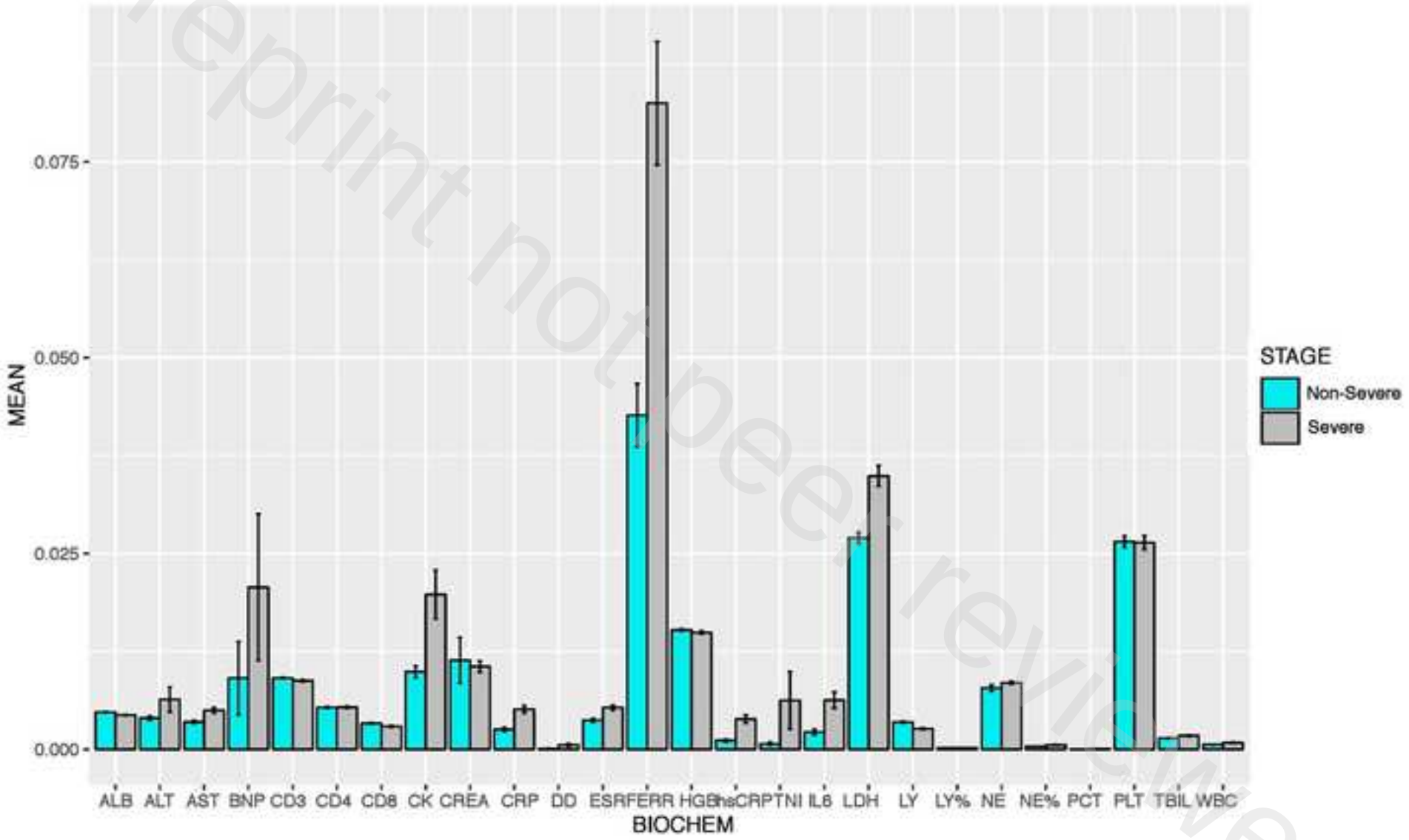
Figure

Click here to access/download;Figure;figure 2.tif



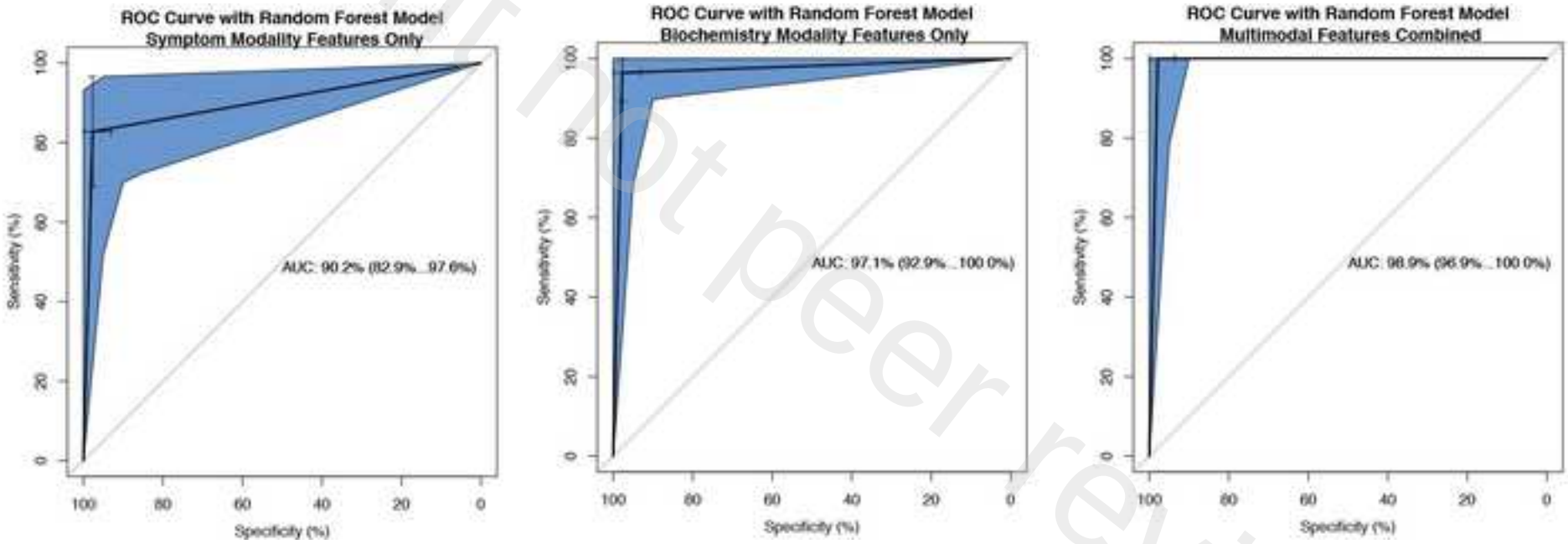
Figure

Click here to access/download;Figure;figure 3.tif



Figure

Click here to access/download;Figure;figure 4.tif



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

Yuanfang Chen[1,2†], Liu Ouyang[3†], Forrest Sheng Bao[4†], Qian Li[5†], Lei Han[2,6], Baoli Zhu[2,6,7], Yaorong Ge[10], Patrick Robinson[8], Ming Xu[2,6,8*], Jie Liu[9*], and Shi Chen[8,11*]

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

† These authors contributed equally to this work.

*** Corresponding to:** Ming Xu, M.D. & Ph.D., Department of Occupational Disease Prevention, Jiangsu Provincial Center for Disease Control and Prevention, No.172 Jiangsu Road, Nanjing 210009, China. Tel/Fax:86-25-85393212, E-mail: sosolou@126.com; **Corresponding to** Jie Liu, M.D. & Ph.D., Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, No 1277 Liberation Avenue, Wuhan 430022, China;

Tel/Fax:86-27-5393212, E-mail: liu_jie0823@163.com; **And corresponding to** Shi Chen, Ph.D., School of Data Science, University of North Carolina Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA. Tel/Fax: 1-704-687-8426, E-mail:schen56@uncc.edu.

Keywords: COVID-19, machine learning, supervised learning, clinical classification
Running title: machine learning for COVID-19 prognosis

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

Introduction

COVID-19 is a pandemic caused by the novel SARS-CoV-2 virus. As of May 30 2020, it had spread through at least 220 countries and regions, causing more than 6 million cases with almost 400 thousand deaths [1](1). It has become the single most severe pandemic in the 21st century, dwarfing other coronavirus-caused 2003 SARS and 2012 MERS epidemics. COVID-19 is especially challenging to the health professionals and general population. Unlike the precedent SARS and MERS epidemics, COVID-19 patients can be either asymptomatic or symptomatic, both of which are demonstrated to be transmissible of the virus with varying degrees [2-5](2-5). In addition, the distinct clinical types, non-severe and severe, require different treatment and care plans [6](6). Current studies are able to differentiate COVID-19 patients from non-patients, but further detecting non-severe or severe types of COVID-19 is not comprehensively explored. Non-severe type patients can be accommodated with relatively less intensive clinical monitoring and intervention, including treating pre-existing comorbidities, preventing healthcare associated infections and other comorbidities [7] (8). In contrast, severe type patients require close monitoring, usually in ICU with more clinicians [6] (6). Therefore, effectively and efficiently classifying COVID-19 clinical types is essential for triage, resource optimization, and care planning for front-line clinicians, healthcare systems, as well as for the patients [6, 8] (6,7).

Currently, non-severe and severe type are classified based on only a few clinical features in China, including shortness of breath, O_2 saturation, and PaO_2 [9](9). Because of the complexity of COVID-19's pathological, physiological, and immunological response, which do not comprehensively characterize the complicated pathology, physiology, and immunological response, these three features do not sufficiently characterize the difference between non-severe and severe types in COVID-19 patients [9-11] (9-11). In addition, some severe patients may not present shortness of breath initially. But without proper medical intervention, their clinical course will worsen abruptly, often resulting in respiratory failure with high mortality [6] (6). Misclassification of COVID-19 clinical types can result in inappropriate early treatment decisions, putting patients at risk of progression due to insufficiently aggressive supportive therapy, or exposing other patients to overly invasive treatment, with negative clinical consequences.

It is therefore critical to provide a rapid, accurate and efficient method to determine severity of COVID-19 infection. Such a determination will allow optimization of treatment plans for

patient care and improve utilization of healthcare resources and staff. We suggest that using additional readily available clinical features, including patient's comorbidities (e.g., hypertension and diabetes), clinical symptoms (e.g., fever and chest pain), and laboratory testing results, are able to develop an effective method to determine COVID-19 clinical type and severity [12, 13](12,13). The human body is a unified and integrated entity. When pathogens such as SARS-CoV-2 invade, its effects can be shown not only from CT scans in the thoracic region, but also from other aspects such as clinical symptoms and laboratory testing results. ACE-2 receptors, which facilitate SARS-CoV-2 infiltration, are distributed across multiple organs and systems in human body [14](35). More recent discoveries have found that in addition to respiratory system, SARS-CoV-2 can also invade digestive, reproductive, and even neural systems as well [15-17](14-17). In other words, comorbidities, clinical symptoms, and laboratory testing information of COVID-19 patients could all be consequences and/or risk factors of SARS-CoV-2 infection. In clinical practice against COVID-19, clinicians not from respiratory or intensive care units may rely only on the referenced symptoms and signs [9](9) while neglecting diverse and important clinical features of COVID-19 patients, and may miss the critical signs of clinical course, leading to undesirable prognosis.

The potential power of symptoms, comorbidities, laboratory testing results, as well as their combinations to determine COVID-19 clinical type is currently not well understood nor evaluated [18-20](18-20). In order to utilize such diverse multimodality clinical information to make accurate and interpretable classifications, we propose a data mining and machine learning (ML) framework alternative to commonly used hypothesis-driven parametric models such as logistic regression. The goal of this study is to provide reliable diagnostic decision support for clinicians even without comprehensive experience on the emerging COVID-19. We aim to explore and contrast the distributions of comorbidities and symptoms, as well as laboratory testing results between non-severe and severe COVID-19 types. We will identify key features that differed substantially between the two clinical types and provide clear evidence-based interpretations for clinicians and other health professionals. Next, we will investigate whether single modality or specific combination of features across modalities are able to provide accurate classification models based on ML techniques. This study delivers an accurate diagnostic decision support tool to differentiate non-severe from severe type patients based on commonly available clinical data with clear clinical interpretations. Insights gained from this study, as well as developed end-to-

end multimodal data analysis and ML framework, will enable us to better understand the comprehensive pathology of COVID-19, further distinguish COVID-19 from other infectious respiratory diseases, and apply in other diseases with multimodal clinical data in the future.

Materials and Methods

Data Source and Clinical Feature Extraction

In this study, we recruited 362 COVID-19 patients from January 2020 to March 2020, including 148 patients presenting severe disease and 214 patients lacking criteria for severe disease during admission from Wuhan Union Hospital, China. Definitions of non-severe and severe cases were mainly adopted from the official COVID-19 Diagnosis and Treatment Plan from the National Health Commission of China and consulted guidelines from American Thoracic Society as well [9-11](9-11). Patients in severe type should present any one of the following features: 1) respiratory rate > 30 breaths per minute; 2) oxygen saturation < 93% at rest; or 3) arterial oxygen partial pressure (PaO_2)/fraction of inspired oxygen (FiO_2) < 300mmHg (40kPa). Each COVID-19 patient was confirmed by two independent qRT-PCR tests before admitted to this study. All patients or their responsible surrogates signed informed consent forms before study entry. Symptoms were evaluated and blood samples were drawn upon admission to perform laboratory testing. No pediatric patients under 18 years were admitted.

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

In addition, we also collected patients' laboratory testing data. Laboratory tests were plasma, serum or whole blood assays for commonly obtained biochemistries, complete blood count with differential counts and percents, immunologic markers, such as IL-6, D-dimer and hsCRP. After initial screening, several features with too many missing data such as calcitonin were excluded. In addition, respiratory rate, oxygen saturation, and PaO_2 without supplemental oxygen were also excluded because they were clinical type defining features according to the Diagnosis and Treatment Plan [9](9). There were 26 laboratory testing features in this study. A detailed description and units of these features were provided in supplementary table 2. All these

laboratory testing features were continuous features, different from the binary features in the symptom modality.

Patient-specific identifying information (e.g., name, address of residence) was removed from data collected for this study. This study was evaluated and approved by the IRB committee of Union Hospital, Wuhan, China (approval number: 2020-IEC-J-345).

Data Mining on Multimodal Clinical Features

Initial data mining on the multimodal COVID-19 data was conducted. Approximately 5% data were missing and predictive mean matching (PMM) was applied to impute the missing data. To evaluate the effectiveness of PMM, we used a subset of the original dataset with no data missing, randomly dropped 5% data to simulate potential data loss, re-extrapolated the data with PMM, and evaluated the mean square root error (RMSE) between the original and imputed datasets. The RMSE was less than 0.05, indicating the extrapolation was feasible and reliable. The imputed data were then passed on to successive data mining and machine learning steps.

The prevalence of each symptom feature was calculated as the number of positives over the number of patients in non-severe and severe groups, defined by the Diagnosis and Treatment Plan [9](9). *Z*-test was applied to detect any statistically significant difference of features between the two types. In addition, a forest plot of odds ratio (OR) and its 95% confidence interval (CI) of symptom features between severe and non-severe COVID-19 types was developed.

For the continuous laboratory testing features, we characterized and contrasted the distribution of each feature between the two types. Feature scaling was performed first to normalize all features between 0 and 1 as laboratory testing values had very different units (Supplementary Table 2). Because most features' values were not normally distributed, we then applied a two-sided Kolmogorov-Smirnov test instead of Student's *t*-test to determine whether distributions of the feature values differed significantly between the two clinical types. Additionally, principal component analysis (PCA) was applied to help visualize the distinction in feature distributions and associations between the two types.

COVID-19 Clinical Type Classification via Machine Learning

Traditional hypothesis-driven parametric models, such as logistic regression, relied heavily on human decisions of how features interact with each other (e.g., interaction terms in logistic regression model), which might not reflect the underlying medical reality. In addition, these models had strict prerequisites to perform correctly, including normality of residuals,

homoscedasticity, and independence of input features. Initial exploratory analyses showed that input features in both symptom and laboratory testing modalities had non-normality and high collinearity among the features. Another technical challenge to logistic regression in this study was a mixture of binary symptom and continuous laboratory testing input features.

Bearing these problems, logistic regression would not be a preferable modeling approach to accurately classify and predict COVID-19 clinical types. Our exploratory analysis showed that logistic regression could only achieve 68% and 77% predictive accuracy on an 80-20 training-prediction split, using symptom and laboratory testing features, respectively (supplementary Table S3). Thus, traditional hypothesis-driven models such as logistic regression were less feasible in clinical settings requiring high accuracy, sensitivity, and specificity to differentiate COVID-19 clinical types and develop personalized treatment plans.

On the other hand, state-of-the-art machine learning (ML) classification models worked directly with the data to avoid human bias. In addition, ML models did not have restrictions on how input data should be distributed or related. Therefore, ML classification would be a more appropriate modeling approach to predict COVID-19 clinical type with complicated data structure in this study. We developed an end-to-end ML framework to accurately predict COVID-19 patient's clinical type based on symptom and/or laboratory testing modality features. We built random forest (RF) classification models, as RF was able to provide excellent interpretability of input variable's relative importance to support clinical decision-making. RF was a widely used ML model based on decision theory and decision tree approach. The internal validation process through bagging made RF especially accurate and reliable. RF was also robust against data loss and data unbalancing, e.g., more non-severe type patients than severe patients in our study [21-25](37-41). There were other types of ML classification methods though, for example, k-nearest neighbor, artificial neural network, and naive Bayes. However, the major goal of this study was not to compare performances of different ML models, we focused on RF to deliver the most accurate classification possible.

We assigned severe cases as "positive" and non-severe as "negative" in the classification. The goal of ML classification through RF was to accurately predict the patient's COVID-19 type, either "positive" (severe) or "negative" (non-severe), based on features from different clinical modalities. In this part of study, we first use a single modality of features as input. The detailed RF modeling and validation process were provided in supplementary material and method. We

trained the model 100 independent runs, each run with a randomly selected set of 80% data for training and the remaining 20% for prediction. This step explored whether RF model was robust against different input data and assessed the model's generalizability. Hyperparameter number of trees (ntree) in the RF model was set at a very low value (ntree=8) to avoid potential overfitting issues [22](38). Important ML performance metrics, including accuracy, sensitivity, specificity, F1 score, and area under curve (AUC) value based on receiver operating characteristic (ROC) curve were computed for the prediction set. In addition, RF was able to evaluate input variables' importance to differentiating the two types based on their Gini impurity [24](40). We further quantified input features' relative importance, identified top contributing features, and explored their clinical relevance and interpretability to COVID-19. The most important features to differentiate COVID-19 clinical types were also cross-checked with our results from exploratory data mining, including prevalence of symptom and distribution of laboratory testing features.

COVID-19 Clinical Type Classification with Multimodal ML

In addition, we explored whether and how combining features across modalities improved classification performance. We developed another RF model that incorporated features from both modalities. The modeling process was exactly the same as using single modality. Nevertheless, instead of putting all features into the model, we selected top 5 features from each of the two modalities as new inputs. These top features were identified from the Gini impurity of single modality RF models (highlighted in supplementary Table S1 and S2). We explored whether a few important features from different modalities could perform sufficiently well to address the clinical challenge of differentiating non-severe and severe COVID-19 patients.

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

Results

COVID-19 Clinical Findings in Non-severe and Severe Disease Types

Prevalence of symptom features in non-severe and severe COVID-19 patients at time of entry into the study were calculated and compared (Fig. 1). Patients in the two clinical types showed distinct prevalence for a number of different features. Severe COVID-19 patients were statistically much more likely to be elderly (at least age of 50, symbol OLD, OR=13.77, 95% CI= 7.33-25.86, $p<0.001$) and male (SEX, OR=1.89, 95% CI= 1.24-2.90, $p<0.01$), to have renal

diseases (KID, OR=8.51, 95% CI= 1.86-38.99, p<0.001), cardiovascular diseases (CAR, OR=5.61, 95% CI=2.81-11.20, p<0.001), hypertension (HYP, OR=5.37, 95% CI=3.36-8.56, p<0.001), diabetes (DIA, OR=4.61, 95% CI=2.53-8.38, p<0.001), loss of appetite and taste (NAP, OR=3.20, 95% CI=1.70-6.01, p<0.001), feeling chilly (CHL, OR=2.21, 95% CI=1.16-4.22, p<0.05), and chest congestion (SHB, OR=1.88, 95% CI=1.22-2.89, p<0.01) than their non-severe counterparts. The only exception was sore throat, where severe patients had significantly much less likelihood to develop (SOR, OR=0.30, 95% CI=0.14-0.61, p<0.001). These discoveries were further demonstrated in the forest plot of odds ratio (OR) and CI in Fig. 2, showing the differences between the two clinical types. Therefore, these relatively easily measured and acquired clinical features could be utilized to clinically evaluate COVID-19 patients' severity. Our findings especially on severe COVID-19 patients echoed the U.S. CDC's recently updated list of symptoms of COVID-19 [26](21) and more recent characterizations of COVID-19 patients in the U.S. [27](36). Our findings showed that elderly male COVID-19 patients with cardiovascular, respiratory, renal diseases and diabetes were at much higher risk of developing serious complications of COVID-19 such as acute respiratory distress syndrome (ARDS) and even death [19, 20](19,20). In addition, we discovered that Chinese patients with renal diseases were significantly more likely to develop severe COVID-19, which was not widely reported before. Clinical evidence showed that ACE-2 expression was associated with kidney diseases, thus making kidney disease a potential complication of SARS-CoV-2 invasion [28, 29](22,23). This finding would inform clinicians to monitor kidney dysfunction, e.g., acute kidney injury, as a clinical sign and/or consequence of severe COVID-19 complication as well.

For laboratory testing modality features, we compared the feature scaled distributions of these continuous features between non-severe and severe types. The results were demonstrated in Fig. 3. Based on the two-sided Kolmogorov-Smirnov test, severe and non-severe COVID-19 types differed significantly in most laboratory features, except platelet (PLT), hemoglobin (HGB), CD3, and CD4. Among all laboratory features, IL-6, hsTNI, and D-dimer were the most clinically important between non-severe and severe COVID-19 types.

In addition, supplementary Fig. 1 showed symptom and laboratory testing PCA results between non-severe and severe types. PCA plots reinforced the conclusion that associations among the features were substantially different between the two clinical types. The two types not

only had vastly different distributions of features, interrelationships among features were also distinct between the two types.

In conclusion, after extensive clinical feature extraction and data mining, there was strong qualitative and quantitative evidence that non-severe and severe COVID-19 types differed substantially with regard to comorbidities, symptoms, and laboratory testing results. These findings paved the way toward creating an effective machine learning (ML) classifier to accurately differentiate these two types in clinical practice.

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

We first explored whether the relatively simple binary features could provide accurate insights in identifying COVID-19 disease severity. Model performance is summarized in Table 1 upper section, using all 26 symptom features as input. Based on 100 independent runs, the RF model reached an average of >99% and 92% accuracy for training and testing sets, respectively (Table 1). AUC was 90.2% (82.9%-97.6%) based on the receiver operating characteristic (ROC) curve (Fig. 4 left panel). The model performed better in detecting true positives (i.e., severe clinical type) than true negatives (i.e., non-severe type). In other words, symptom features alone in RF models were very unlikely to misclassify severe case as non-severe case, but with a higher chance to predict non-severe case to severe case. In clinical practice, this would be a lesser concern, as false positive (failed to detect non-severe type) would be more tolerable than false negative (failed to detect severe type).

Our RF model also identified the major influential features to differentiate COVID-19 types based on contribution to Gini impurity. Top influential features in descending order were age, gender, hypertension, diabetes, and cardiovascular diseases, in accordance with existing literature [30](24). Other important symptom features included fatigue, chest congestion, sore throat, phlegm production, and fever. Most of these findings aligned well with our parametric data mining with odds ratio (OR) comparison (Fig. 2, supplementary Table S1) but with much higher accuracy (90% accuracy on prediction set of RF model compared to 68% accuracy from non-ML logistic regression). The only exception was renal disease. While its prevalence was significantly different between the two types, the RF model did not consider it as a major differentiating factor based on Gini impurity (Supplementary Table S1). Clinically, elderly male patients with pre-existing comorbidities, especially hypertension, diabetes, cardiovascular

diseases were much more vulnerable to COVID-19 and had a much higher risk to develop to severe type [18, 20](18,20). Therefore, we suggested using COVID-19 patients' demographic, comorbidity and symptom features as the first round of evaluation of severity with reasonable accuracy.

Clinical Type Classification via ML: Laboratory Testing Modality

Using RF model with 26 laboratory testing features was even more effective in differentiating non-severe and severe types. On average of 100 independent runs, the RF model achieved >99% and >95% accuracy for training and testing sets, respectively. Sensitivity, specificity, and F1 scores were all above 95%, using only 8 trees in the RF model (Table 1, middle section). AUC was 98% based on ROC curve (Fig. 4 middle panel). Though this study focused on ML methods, we evaluated model performance of non-ML logistic regression (<80% accuracy) in supplementary Table S3 as a reference point to show the improvement that state-of-the-art ML models could achieve.

Top differentiating features in laboratory testing modality were D-dimer (DD), high sensitivity troponin I (hsTNI), absolute neutrophil counts (NE), interleukin-6 (IL-6), lactate dehydrogenase (LDH), and high sensitivity c-reactive protein (hsCRP), in descending order. The clinical interpretation of their important role was that severe COVID-19 patients had more intensive immune response and hyperinflammation, such as cytokine storm syndrome with substantially increased IL-6 [31](25). Research also showed that SARS-CoV-2 was able to infect many organs other than lungs and induce dysfunction of these organs, including heart [32, 33](26,27). Increasing hsTNI was a sign of heart tissue damage from SARS-CoV-2 infection [34](28). In addition, severe COVID-19 patients might have formed microthrombosis which induced higher D-dimer [18, 20, 35-37](18, 20,29-31). Abnormal level of neutrophils could be responsible for cytokine storm and ARDS in severe COVID-19 patients [13, 38](13,32). hsCRP, a biomarker of acute inflammation, cardiovascular disease, and ischemic events, was also confirmed as the major contributing factor of COVID-19 mortality [18](18). LDH was a biomarker of tissue damage and was used to predict the clinical course of COVID-19 patients [39](42). These findings added further clinical insights in how multiple organs and systems, not just lungs, responded to SARS-CoV-2 infection in different clinical types [14, 40, 41](33-35).

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

Clinical Type Classification via ML: Multimodal Features

We further developed a multimodal RF model that incorporated both symptom and laboratory testing modalities. We used only the 5 most influential features from symptom and 5 most influential features from laboratory testing modalities, based on their Gini impurity values. The results showed that the top 10 out of a total of 52 features from both modalities achieved >99% in every model performance metric, including accuracy, sensitivity, specificity, and F1 score (Table 1, 3rd section). AUC was >99% as well (Fig. 4 right panel).

These findings reinforced our argument that SARS-CoV-2 attacked multiple organs and systems, and the human body reacted in a unity against its invasion. Different clinical features (e.g., comorbidity, symptom, and laboratory testing results) complemented each other to provide a more comprehensive characterization of human body as a united entity, not just respiratory system, reacted to SARS-CoV-2 invasion [14](35). In addition, the decent model performance promised the feasibility of multimodal clinical data mining in detecting and differentiating non-severe from severe COVID-19 patients. Our work would help effectively optimize healthcare operation during the pandemic and avoid overloading the healthcare system [8](7).

Discussion

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

In addition to accurately detecting vulnerable COVID-19 patients who are likely to be in severe type, this study also provides clinical insights on why these patients may have been in severe type. Machine learning (ML) models work directly with data and therefore are generally not good at providing clear interpretations. In this study, we combine the power of both hypothesis-driven and data-driven ML models. The most contributing comorbidity, symptom,

and biochemical features help predict and explain potential COVID-19 clinical courses and prognosis. Our research echoes recent studies that characterize and predict clinical course, critical illness and mortality of COVID-19 patients [13, 18, 20](13,18,20). In particular, another decision tree-based algorithm (XGBoost) showed promising performance in predicting mortality of CoVID-19 patients [18](18). RF was technically similar to XGBoost and our results were consistent to identify the key differentiating features, including LDH and hsCRP.

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

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

The next step of this study is to further include biomedical imaging modality. A technical barrier is that CT scan is a high-dimensional feature set while symptom and laboratory testing have relatively low dimensionality. Therefore, CT scan, at its original form of imaging, cannot be effectively combined with other modalities. We will evaluate the feasibility of using convolution neural network (CNN, another time of ML technique) first to reduce feature space in

CT scan and extracting a fully connected layer in CNN as a representation of CT scan feature. A fully connected layer is a 1-dimensional vector and has the same dimensionality with the other two modalities. Therefore, in theory we would be able to further combine CT scans with other clinical features and investigate the association between these features with regard to COVID-19.

COVID-19 is a complex disease where the pathogen not only attacks the respiratory system but other organs and systems that have ACE-2 receptors as well [14, 27](35,36). Our findings reveal the complicated pathological, physiological, and immunological responses to SARS-CoV-2 invasion and shed light in understanding the complex interactions between the virus and human body. Though the multimodal data mining and ML framework is developed with severe/non-severe COVID-19 data, we suggest that the end-to-end framework is applicable to many disease systems where multimodal inputs are common, including demographic information, comorbidity, laboratory testing, imaging, and -omics data. Having a more holistic viewpoint and approach will enable us to understand and respond to these emerging diseases, especially the unprecedented COVID-19, more readily in the field. We will further explore this analytical framework, and transfer insights for future clinical studies such as differentiating healthy, non-COVID viral pneumonia, non-severe, and severe COVID-19 patients.

In this study, we recruited participants from a single hospital in Wuhan, the first epicenter of COVID-19. There will inevitably be selection bias, as currently the ethnicity group is limited to Chinese participants. Therefore, we want to inform our colleagues across the world and see whether different demographic backgrounds influence feature distributions between non-severe and severe types in the patients. Our findings have already been independently identified in COVID-19 patients across ethnicity groups [26, 34](21, 28). Another pitfall we should be aware of is that comorbidities may be the consequence of SARS-CoV-2 invasion, or risk factors that increase the risk of infection. Although the participants' comorbidity, symptom, and laboratory testing were evaluated upon admission to hospital, they could have been exposed to the pathogen long before hospitalization, given the long "asymptomatic" type of COVID-19. For example, it is unclear whether kidney damage in a patient is a risk factor to induce severe type COVID-19, or SARS-CoV-2 attacks the kidney and causes kidney damage [14] (35). The causal relationship needs to be more systematically evaluated with carefully designed prospective cohort studies. Nevertheless, in clinical practice, observing renal diseases in COVID-19 patients would trigger

an alarm of clinical course to severe type, and inform clinicians to take actions to prevent acute kidney failure and even death.

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

Conclusion

We trained and validated machine learning random forest (RF) models to predict COVID-19 severity based on 26 comorbidity and symptom features and 26 laboratory testing features from a cohort of 214 non-severe and 148 severe type COVID-19 patients. We identified top features from both feature modalities to differentiate clinical types, and achieved predictive accuracy of >90%, >95%, and >99% when comorbidity/symptom, laboratory data, and top features from each modality combined were used as input, respectively. The results will help clinicians evaluate COVID-19 patient's severity and triage more effectively, and optimize healthcare resource utilization during this pandemic.

Acknowledgments

This work is a tribute to the frontline clinicians and supporting staff who devoted their lives to combating this COVID-19 pandemic. This study was jointly supported by the National Science Foundation for Young Scientists of China (81703201), the Natural Science Foundation for Young Scientists of Jiangsu Province (BK20171076), the Jiangsu Provincial Medical Innovation Team (CXTDA2017029), the Jiangsu Provincial Medical Youth Talent program (QNRC2016548), the Jiangsu Preventive Medicine Association program (Y2018086), the Lifting Program of Jiangsu Provincial Scientific and Technological Association, and the Jiangsu Government Scholarship for Overseas Studies.

Author contributions

Y.C. designed the study. L.O. and J.L. derived and processed data. L.O., F.S.B., Q.L., L.H., B.Z, J.L., P.R. and S.C. interpreted results. M.X. and S.C. performed analyses. Y.G. and S.C. supervised. P.R. and S.C. developed the manuscript. Y.C., L.O., and F.S.B. contributed equally.

Competing interests

The authors declare no competing interests in this study.

Data and materials availability

De-identified clinical data and codes in this study are openly available on GitHub (https://github.com/forrestbao/corona/tree/master/blood).

References:

1. World Health Organization. Coronavirus Disease 2019 (COVID-19) Situation Report 118. *https://wwwwhoint/docs/default-source/coronaviruse/situation-reports/20200517-covid-19-sitrep-118pdf?sfvrsn=21c0dafe_6* 2020
2. Gandhi M, Yokoe DS, Havlir DV. Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19. *The New England journal of medicine* 2020; **382**(22): 2158-60.
3. Wolfel R, Corman VM, Guggemos W, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature* 2020; **581**(7809): 465-9.
4. Luo Y, Trevathan E, Qian Z, et al. Asymptomatic SARS-CoV-2 Infection in Household Contacts of a Healthcare Provider, Wuhan, China. *Emerging infectious diseases* 2020; **26**(8).
5. Ye F, Xu S, Rong Z, et al. Delivery of infection from asymptomatic carriers of COVID-19 in a familial cluster. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* 2020; **94**: 133-8.
6. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. *Jama* 2020.
7. Chen S, Zhang Z, Yang J, et al. Fangcang shelter hospitals: a novel concept for responding to public health emergencies. *Lancet* 2020; **395**(10232): 1305-14.
8. Truog RD, Mitchell C, Daley GQ. The Toughest Triage - Allocating Ventilators in a Pandemic. *The New England journal of medicine* 2020; **382**(21): 1973-5.
9. National Health Commission of China. Novel coronavirus pneumonia diagnosis and treatment plan (provisional 7th edition). 2020.
10. Metlay JP, Waterer GW, Long AC, et al. Diagnosis and Treatment of Adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America. *American journal of respiratory and critical care medicine* 2019; **200**(7): e45-e67.
11. Neill AM, Martin IR, Weir R, et al. Community acquired pneumonia: aetiology and usefulness of severity criteria on admission. *Thorax* 1996; **51**(10): 1010-6.
12. Epidemiology Working Group for Ncip Epidemic Response CCfDC, Prevention. [The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)

in China]. *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi* 2020; **41**(2): 145-51.
13. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020; **395**(10223): 497-506.
14. Wadman M, Couzin-Frankel J, Kaiser J, Matacic C. How does coronavirus kill? Clinicians trace a ferocious rampage through the body, from brain to toes. *Science* 2020.
15. Cholankeril G, Podboy A, Aivaliotis VI, et al. High Prevalence of Concurrent Gastrointestinal Manifestations in Patients with SARS-CoV-2: Early Experience from California. *Gastroenterology* 2020.
16. Jing Y, Run-Qian L, Hao-Ran W, et al. Potential influence of COVID-19/ACE2 on the female reproductive system. *Molecular human reproduction* 2020.
17. Moriguchi T, Harii N, Goto J, et al. A first case of meningitis/encephalitis associated with SARS-Coronavirus-2. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* 2020; **94**: 55-8.
18. Yan L, Zhang H, Goncalves J, et al. An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020; **2**: 283-8.
19. Liu J, Li S, Liu J, et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine* 2020; **55**: 102763.
20. Liang W, Liang H, Ou L, et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. *JAMA internal medicine* 2020.
21. J.R. Q. Introduction to decision tree. *Machine Learning* 1986; **1**: 81-106.
22. Ho TK. Random Decision Forest. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal* 1995: 278-82.
23. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence volume* 2020; **2**: 56-67.
24. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* 2011; **27**(14): 1986-94.
25. Strobl C, Boulesteix A-L, Augustin T. Unbiased split selection for classification trees based on the Gini index. *Computational Statistics & Data Analysis* 2007; **52**(483-501).
26. Centers for Disease Control and Prevention. Symptoms of Coronavirus (COVID-19). *https://wwwcdcgov/coronavirus/2019-ncov/symptoms-testing/symptomshtml* 2020.
27. Garg S, Kim L, Whitaker M, et al. Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 - COVID-NET, 14 States, March 1-30, 2020. *MMWR Morbidity and mortality weekly report* 2020; **69**(15): 458-64.
28. Mizuiri S, Ohashi Y. ACE and ACE2 in kidney disease. *World journal of nephrology* 2015; **4**(1): 74-82.
29. Lieben L. Alport syndrome: ACE2 administration slows kidney damage. *Nature reviews Nephrology* 2017; **13**(5): 261.
30. Yang X, Yu Y, Xu J, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory medicine* 2020; **8**(5): 475-81.
31. Li Y, Hu Y, Yu J, Ma T. Retrospective analysis of laboratory testing in 54 patients with severe- or critical-type 2019 novel coronavirus pneumonia. *Laboratory investigation; a journal of technical methods and pathology* 2020; **100**(6): 794-800.

32. Chen L, Li X, Chen M, Feng Y, Xiong C. The ACE2 expression in human heart indicates new potential mechanism of heart injury among patients infected with SARS-CoV-2. *Cardiovascular research* 2020; **116**(6): 1097-100.
33. Shi S, Qin M, Shen B, et al. Association of Cardiac Injury With Mortality in Hospitalized Patients With COVID-19 in Wuhan, China. *JAMA cardiology* 2020.
34. Li JW, Han TW, Woodward M, et al. The impact of 2019 novel coronavirus on heart injury: A Systematic review and Meta-analysis. *Progress in cardiovascular diseases* 2020.
35. Wu C, Chen X, Cai Y, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA internal medicine* 2020.
36. Guan WJ, Ni ZY, Hu Y, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *The New England journal of medicine* 2020; **382**(18): 1708-20.
37. Griffin DO, Jensen A, Khan M, et al. Pulmonary Embolism and Increased Levels of d-Dimer in Patients with Coronavirus Disease. *Emerging infectious diseases* 2020; **26**(8).
38. Barnes BJ, Adrover JM, Baxter-Stoltzfus A, et al. Targeting potential drivers of COVID-19: Neutrophil extracellular traps. *The Journal of experimental medicine* 2020; **217**(6).
39. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020; **395**(10229): 1054-62.
40. Zaim S, Chong J, Sankaranarayanan V, Harky A. COVID-19 and Multi-Organ Response. *Cardiology* 2020: pii:S0146-2806(20)30095-5.
41. Wang T, Du Z, Zhu F, et al. Comorbidities and multi-organ injuries in the treatment of COVID-19. *Lancet* 2020; **395**(10228): e52.

Figure Legends:

Fig. 1. Symptom Features Comparison between Non-severe and Severe Stage COVID-19 Patients.
Note: symptom features were binary, so Y-axis was the prevalence of positives.

Fig. 2. Forest Plot of Top Symptom Features that Differ Significantly between Stages.
Note: *** $p<0.001$; ** $p<0.01$; * $p<0.05$ from the 2x2 contingency table. Forest plot is based on parametric statistical analysis and is irrelevant to random forest, a type of machine learning model used later in this study. The threshold for a feature to be "positively" or "negatively" associated with severe COVID-19 was 1 (dashed line), not 0.

Fig. 3. Laboratory Testing Features Comparison between Non-severe and Severe Stage COVID-19 Patients.
Note: values shown on y-axis were after feature scaling and were between 0 and 1. Error bars represented standard error (SE) of each laboratory testing feature.

Fig. 4. ROC Curve from Random Forest Model Based on Symptom, Laboratory Testing, and Multimodal Features.
Note: panel (A) left showed symptom feature as input alone, panel (B) middle showed laboratory testing as input alone, and panel (C) right showed both features combined as input.

Table 1. Random Forest Model Prediction Performance with Multimodal Features

Feature	Median (%)	Minimum (%)	Maximum (%)
Symptom			
Accuracy%	90.28	83.33	98.61
Sensitivity%	97.5	87.80	> 99
Specificity%	81.48	60.00	> 99
F1 Score%	93.31	84.62	98.97
AUC%	90.20	82.90	97.60
Laboratory Testings			
Accuracy%	97.22	94.44	> 99
Sensitivity%	97.92	91.11	> 99
Specificity%	96.97	87.5	> 99
F1 Score%	97.89	95.35	> 99
AUC%	97.10	92.90	> 99
Multimodal			
Accuracy%	> 99	97.22	> 99
Sensitivity%	> 99	97.22	> 99
Specificity%	> 99	92.00	> 99
F1 Score%	> 99	97.22	> 99
AUC%	98.90	96.90	> 99

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