Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study

Juan Carlos Quiroz1,2*, PhD; You-Zhen Feng3, MD; Zhong-Yuan Cheng3*, MD; Dana Rezazadegan1,4, PhD; Ping-Kang Chen3, MD; Qi-Ting Lin3, MD; Long Qian5, PhD; Xiao-Fang Liu6,7, PhD; Shlomo Berkovsky1, PhD; Enrico Coiera1, PhD; Lei Song8, MD; Xiaoming Qiu9, MD; Sidong Liu1, PhD; Xiang-Ran Cai3, MD

1Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine, Health and Human Sciences, Macquarie University, Macquarie Park, Australia
2Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia
3Medical Imaging Centre, The First Affiliated Hospital of Jinan University, Guangzhou, China
4Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne, Australia
5Department of Biomedical Engineering, Peking University, Beijing, China
6Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Nankai University, Tianjin, China
7School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
8Department of Radiology, Xiangyang Central Hospital, Affiliated Hospital of Hubei University of Arts and Science, Xiangyang, China
9Department of Radiology, Huangshi Central Hospital, Affiliated Hospital of Hubei Polytechnic University, Edong Healthcare Group, Huangshi, China
*these authors contributed equally

Corresponding Author:
Sidong Liu, PhD
Centre for Health Informatics, Australian Institute of Health Innovation
Faculty of Medicine, Health and Human Sciences
Macquarie University
75 Talvera Road
Macquarie Park, 2113
Australia
Phone: 61 29852729
Email: sidong.liu@mq.edu.au

Abstract

Background: COVID-19 has overwhelmed health systems worldwide. It is important to identify severe cases as early as possible, such that resources can be mobilized and treatment can be escalated.

Objective: This study aims to develop a machine learning approach for automated severity assessment of COVID-19 based on clinical and imaging data.

Methods: Clinical data—including demographics, signs, symptoms, comorbidities, and blood test results—and chest computed tomography scans of 346 patients from 2 hospitals in the Hubei Province, China, were used to develop machine learning models for automated severity assessment in diagnosed COVID-19 cases. We compared the predictive power of the clinical and imaging data from multiple machine learning models and further explored the use of four oversampling methods to address the imbalanced classification issue. Features with the highest predictive power were identified using the Shapley Additive Explanations framework.

Results: Imaging features had the strongest impact on the model output, while a combination of clinical and imaging features yielded the best performance overall. The identified predictive features were consistent with those reported previously. Although oversampling yielded mixed results, it achieved the best model performance in our study. Logistic regression models differentiating between mild and severe cases achieved the best performance for clinical features (area under the curve [AUC] 0.848; sensitivity 0.455; specificity 0.906), imaging features (AUC 0.926; sensitivity 0.818; specificity 0.901), and a combination of clinical and imaging features (AUC 0.950; sensitivity 0.764; specificity 0.919). The synthetic minority oversampling method further improved the performance of the model using combined features (AUC 0.960; sensitivity 0.845; specificity 0.929).
Conclusions: Clinical and imaging features can be used for automated severity assessment of COVID-19 and can potentially help triage patients with COVID-19 and prioritize care delivery to those at a higher risk of severe disease.

(JMIR Med Inform 2021;9(2):e24572) doi: 10.2196/24572

KEYWORDS
algorithm; clinical data; clinical features; COVID-19; CT scans; development; imaging; imbalanced data; machine learning; oversampling; severity assessment; validation

Introduction
COVID-19 has overwhelmed health systems worldwide [1,2]. Considering the various complications associated with COVID-19 [3-5], methods that help triage patients with COVID-19 can help prioritize care delivery to individuals at a high risk of severe or critical illness. COVID-19 severity can be categorized as follows: mild, ordinary, severe, and critical [6]. Severe and critical cases require intensive care and more health care resources than mild and ordinary cases. A high rate of false-positive severe or critical cases could overwhelm health care resources (ie, beds in the intensive care unit). Moreover, delays in identifying severe or critical cases would lead to delayed treatment of patients at a higher risk of mortality. Therefore, it is important to identify severe cases as early as possible, such that resources can be mobilized and treatment can be escalated.

Chest computed tomography (CT) scans provide important diagnostic and prognostic information [7,8]; consequently, they have been the focus of numerous recent studies using machine learning techniques for COVID-19–related prediction tasks [9-21]. Previous studies have focused on mortality predictions [9], diagnosis (identifying COVID-19 cases and differentiating them from other pulmonary diseases or no disease) [10-15,19,22-25], and severity assessment and disease progression [16-18,23]. Most current approaches have used deep learning methods and imaging features from CT scans [10-15,19,22-24] and X-ray imaging [18,20,21] with popular architectures including ResNet [10,12,14,23], U-Net [11,17], Inception [15,22], Darknet [20], and other convolutional neural networks (NNs) [18,21,26,27]. Recent reviews provide more details regarding these architectures [1,28-32].

Although automated assessment of chest CT scans to predict COVID-19 severity is of great clinical importance, few studies have focused on it [16-18,23]. Automated assessment of chest CT scans can substantially reduce the image reading time for radiologists, provide quantitative data that can be compared across patients and time points, and can be clinically applicable in disease detection and diagnosis, progression tracking, and prognosis [8]. While CT scans are an important diagnostic tool, previous studies reported that clinical data, such as symptoms, comorbidities, and laboratory findings, differed between patients with COVID-19 admitted to intensive care units and those who were not [33], and these data help predict the mortality risk [9]. A previous study compared the imaging data and clinical data of 81 patients with confirmed COVID-19 and suggested that the combination of imaging features with clinical and laboratory findings facilitated an early diagnosis of COVID-19 [34].

In this study, we used patient clinical data and imaging data to predict disease severity among patients with COVID-19. Considering this as a putative binary classification task, we predicted whether a patient diagnosed with COVID-19 is likely to have mild or severe disease. This study has 3 objectives. First, we compared the predictive power of clinical and imaging data for disease severity assessment by testing three machine learning models: logistic regression (LR) [35], gradient boosted trees (eg, XGBoost) [36], and NNs [37]. Second, since our cohort data are highly imbalanced, with the majority of cases being of mild/ordinary severity, we tested 4 oversampling methods to address the imbalanced classification issue [38-41]. Third, we interpreted the importance of features by using the Shapley Additive Explanations (SHAP) framework and identified features with the highest predictive power [42]. The predictive models evaluated herein yielded high accuracy and identified predictive imaging and clinical features consistent with those reported previously.

Methods
Participants
This retrospective study was performed using data collected by 2 hospitals in the Hubei Province, China. The study cohort consisted of patients with COVID-19 diagnosed through RT–PCR analysis of nasopharyngeal swab samples. A total of 346 patients from 2 hospitals were retrospectively enrolled, including 230 (66.5%) patients from Huang Shi Central Hospital (HSCH) and 116 (33.5%) from Xiang Yang Central Hospital (XYCH). These patients were admitted to hospital between January 1 and February 23, 2020, and underwent chest CT upon initial hospitalization. All participants provided written consent. This study was approved by the institutional review board of both hospitals (approval number LL-2020-032-02). Table 1 summarizes the demographic characteristics of the patients in the 2 cohorts.
Table 1. Demographic characteristics of the patients in the 2 cohorts (N=346).

<table>
<thead>
<tr>
<th>Category</th>
<th>Variables</th>
<th>HSCH(^a) (n=230)</th>
<th>XYCH(^b) (n=116)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COVID-19 severity, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td></td>
<td>7 (3.0)</td>
<td>1 (0.9)</td>
<td>8 (2.3)</td>
</tr>
<tr>
<td>Ordinary</td>
<td></td>
<td>212 (92.2)</td>
<td>104 (89.7)</td>
<td>316 (91.3)</td>
</tr>
<tr>
<td>Severe</td>
<td></td>
<td>7 (3.0)</td>
<td>6 (5.2)</td>
<td>13 (3.8)</td>
</tr>
<tr>
<td>Critical</td>
<td></td>
<td>4 (1.7)</td>
<td>5 (4.3)</td>
<td>9 (2.6)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td></td>
<td>49.0 (14.4)</td>
<td>47.5 (17.2)</td>
<td>48.5 (15.4)</td>
</tr>
<tr>
<td>Gender ratio (female to male)</td>
<td></td>
<td>120:110</td>
<td>57:59</td>
<td>177:169</td>
</tr>
</tbody>
</table>

\(^a\)HSCH: Huang Shi Central Hospital.  
\(^b\)XYCH: Xiang Yang Central Hospital.

**Imaging and Clinical Data**

Chest CT scans of patients were collected upon initial hospitalization and preprocessed using intensity normalization, contrast limited adaptive histogram equalization, and gamma adjustment, using the same preprocessing pipeline as in our previous study [43]. We performed lung segmentation in the chest CT images by using an established model “R231CovidWeb” [44], which was pretrained using a large, diverse data set of non–COVID-19 chest CT scans and further fine-tuned with an additional COVID-19 data set [45]. CT slices with <3 mm\(^2\) of lung tissue were excluded from the data sets since they provide limited or no information about the lung. Lung lesions were segmented using EfficientNetB7 U-Net [16], which was also pretrained using a public COVID-19 data set [45]. The model indicated four types of lesions: ground-glass opacities, consolidations, pleural effusions, and other abnormalities. The volume of each lesion type and the total lesion volume were calculated from the segmentation maps as the imaging features and were further normalized by the lung volume. Figure 1 shows representative results of lung and lesion segmentation of a mild case and a severe case, wherein the upper row presents 3D models of the lung and lesions reconstructed using 3D Slicer (v4.6.2) [46], and the lower row presents axial chest CT slices with the lung and lesion (green: ground-glass opacities, yellow: consolidation, and brown: pleural effusion) boundaries overlaid on the CT slices.
Clinical data collected from the patients included demographic characteristics, signs, symptoms, comorbidities, and the following 18 laboratory findings: white blood cell count ($\times 10^9$/L), neutrophil count ($\times 10^9$/L), lymphocyte count ($\times 10^9$/L), hemoglobin (g/L), platelets ($\times 10^9$/L), prothrombin time (s), activated partial thromboplastin time (s), D-dimer (nmol/L), C-reactive protein (mg/L), albumin (g/L), alanine aminotransferase ($\mu$kat/L), aspartate aminotransferase ($\mu$kat/L), total bilirubin ($\mu$mol/L), potassium (mmol/L), sodium (mmol/L), creatinine ($\mu$mol/L), creatine kinase ($\mu$kat/L), and lactate dehydrogenase ($\mu$kat/L).

All features were either continuous or binary—all binary features include signs, symptoms, and comorbidities. Continuous features were standardized to be centered around 0 (SD 1). Figure 2 shows the structure and dimensions of the features used in this study. These features were grouped into four feature sets: demographic characteristics and symptoms (a subset of the available clinical features), clinical features (demographic characteristics, signs and symptoms, and laboratory findings), imaging features extracted from the chest CT scans through deep learning methods, and a combination of clinical and imaging features.
Severity Assessment Models

We trained and compared three models to predict case severity: LR (with scikit-learn) [47], gradient boosted trees (XGBoost) [36], and an NN (fast.ai) [48]. We used the HSCH data (230 samples) for training and validation using 5-fold repeated stratified cross-validation. The XYCH data (116 samples) were withheld for testing. We reported the results for the test set with the area under the curve (AUC) and F1 scores averaged through independent runs.

Hyperparameter exploration and tuning were performed using the training/validation set. A random search was performed to tune the hyperparameters of LR and XGBoost. For NN, we used a 4-layer, fully connected architecture, with the first hidden layer having 200 nodes and a second hidden layer of 100 nodes. We determined the learning rate (0.01) using Learning Rate Finder [49]. All other NN parameters were set to default values. We explored a different number of nodes in the first and second hidden layers, with 200x100 images yielding the best results in the validation set. Of 346 patients, 167 (48%) had at least one missing feature (5.7 on average, mostly for the laboratory findings). Missing feature values were imputed with the mean for each feature.

Oversampling

The majority of cases in our data set were of mild/ordinary severity, with only a few cases of severe/critical severity. The imbalance ratio for the entire data set was 0.07; training/validation set, 0.05; and testing set, 0.10. We tested four oversampling methods to increase the ratio of the minority class: synthetic minority oversampling (SMOTE) [38], Adaptive Synthetic sampling [39], geometric SMOTE [40], and a conditional generative adversarial network (CTGAN) model for tabular data [41]. For these methods, we oversampled the training set, trained a model using the oversampled data, and reported results on the same test set. We adjusted the resampling ratio of all methods to 0.3 (thus setting the imbalance ratio to 0.3). Using CTGAN for oversampling, we fitted the CTGAN model with the training set, performed sampling to generate synthetic data, using only synthetic data for the minority class (severe/critical), and this was repeated until the minority-to-majority class ratio approached 0.3.

Results

Patient Characteristics

Table 2 summarizes the patients’ characteristics. The differences between the mild/ordinary and severe/critical groups were assessed with the Mann-Whitney U test and Fisher exact test. The median age of the entire cohort was 49 (IQR 38-59) years. The median age of patients with mild/ordinary COVID-19 was 48.5 (IQR 37.0-57.3) years and that of patients with severe/critical COVID-19 was 63.0 (IQR 52.5-69.5) years. We observed significant differences between patients with severe/critical COVID-19 and those with mild/normal COVID-19 with respect to age ($P < .001$) and comorbidities including cardiovascular disease ($P = .002$), hypertension ($P = .002$), diabetes ($P = .01$), and cancer ($P = .01$). From among all signs and symptoms, an increased respiration rate ($P = .002$) and dyspnea ($P < .001$) were more common among patients with severe/critical COVID-19 than among those with mild/ordinary COVID-19.
Table 2. Demographics and baseline characteristics of patients with confirmed COVID-19 (N=346). Symptoms including cardiovascular disease and shortness of breath were more likely in cases of severe/critical COVID-19.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Patients</th>
<th>Mild/ordinary</th>
<th>Severe/critical</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size, n</td>
<td></td>
<td>324</td>
<td>22</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Demographic characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>49.0 (38.0-59.0)</td>
<td>48.5 (37.0-57.3)</td>
<td>63.0 (52.5-69.5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td></td>
<td>.38</td>
</tr>
<tr>
<td>Female</td>
<td>177 (51.2)</td>
<td>168 (51.9)</td>
<td>9 (41.0)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>169 (48.8)</td>
<td>156 (48.1)</td>
<td>13 (59.0)</td>
<td></td>
</tr>
<tr>
<td><strong>Comorbidities, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>40 (11.6)</td>
<td>32 (9.9)</td>
<td>8 (36.0)</td>
<td>.002</td>
</tr>
<tr>
<td>Diabetes</td>
<td>34 (9.8)</td>
<td>28 (8.6)</td>
<td>6 (27.0)</td>
<td>.01</td>
</tr>
<tr>
<td>Hypertension</td>
<td>51 (14.7)</td>
<td>42 (13.0)</td>
<td>9 (41.0)</td>
<td>.002</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>11 (3.2)</td>
<td>9 (2.8)</td>
<td>2 (9.0)</td>
<td>.15</td>
</tr>
<tr>
<td>Chronic liver disease</td>
<td>7 (2.0)</td>
<td>7 (2.2)</td>
<td>0 (0)</td>
<td>N/A</td>
</tr>
<tr>
<td>Chronic kidney disease</td>
<td>4 (1.2)</td>
<td>3 (0.9)</td>
<td>1 (5.0)</td>
<td>.20</td>
</tr>
<tr>
<td>Cancer</td>
<td>8 (2.3)</td>
<td>5 (1.5)</td>
<td>3 (14.0)</td>
<td>.01</td>
</tr>
<tr>
<td><strong>Signs, median (IQR)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body temperature</td>
<td>37.8 (37-38.3)</td>
<td>37.8 (37-38.3)</td>
<td>38.1 (37.1-39)</td>
<td>.11</td>
</tr>
<tr>
<td>Heart rate</td>
<td>85 (80-90)</td>
<td>85 (80-90)</td>
<td>90 (80-101.8)</td>
<td>.11</td>
</tr>
<tr>
<td>Breaths per minute</td>
<td>20 (20-21)</td>
<td>20 (20-21)</td>
<td>21 (20-28)</td>
<td>.002</td>
</tr>
<tr>
<td>Blood pressure high</td>
<td>120 (119.5-130.0)</td>
<td>120 (118.5-130.0)</td>
<td>127 (120-146.5)</td>
<td>.07</td>
</tr>
<tr>
<td>Blood pressure low</td>
<td>74 (69-80)</td>
<td>74 (69-80)</td>
<td>79.5 (71-89)</td>
<td>.08</td>
</tr>
<tr>
<td><strong>Symptoms, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fever</td>
<td>275 (79.5)</td>
<td>256 (79.0)</td>
<td>19 (86.0)</td>
<td>.59</td>
</tr>
<tr>
<td>Cough</td>
<td>238 (68.8)</td>
<td>220 (67.9)</td>
<td>18 (82.0)</td>
<td>.24</td>
</tr>
<tr>
<td>Fatigue</td>
<td>118 (34.1)</td>
<td>108 (33.3)</td>
<td>10 (45.0)</td>
<td>.25</td>
</tr>
<tr>
<td>Dyspnea</td>
<td>32 (9.2)</td>
<td>23 (7.1)</td>
<td>9 (41.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sore muscle</td>
<td>38 (11.0)</td>
<td>35 (10.8)</td>
<td>3 (14.0)</td>
<td>.72</td>
</tr>
<tr>
<td>Headache</td>
<td>34 (9.9)</td>
<td>31 (9.6)</td>
<td>3 (14.0)</td>
<td>.47</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>23 (6.6)</td>
<td>20 (6.2)</td>
<td>3 (14.0)</td>
<td>.17</td>
</tr>
<tr>
<td>Nausea</td>
<td>9 (2.6)</td>
<td>7 (2.2)</td>
<td>2 (9.0)</td>
<td>.11</td>
</tr>
</tbody>
</table>

<sup>a</sup>P values were compared using mild/ordinary and severe/critical cases were obtained with Mann-Whitney U test and Fisher exact test. As no patient in our cohort had a stomach ache, this feature was not factored into our model.

<sup>b</sup>N/A: not applicable.

**Prediction of COVID-19 Severity at Baseline**

Data from the HSCH (230 patients, 66.5%) were used for training and validation, and data from the XYCH (116 patients, 33.5%) were used as the independent test set. We compared model performance using four feature sets: demographic characteristics and symptoms, clinical features, imaging features, and a combination of clinical and imaging features (Figure 2). The optimal classification threshold for the sensitivity, specificity, and F1 score was identified using the Youden index [50]. Table 3 shows the severity assessment performance of an LR model, an XGBoost model, and a 4-layer fully connected NN model. Overall, LR models outperformed the other evaluated models, achieving the highest AUC, F1 score, and sensitivity for all four feature sets. While imaging features yielded substantially better results than clinical features, the combination of clinical and imaging features benefited only the LR model. Hence, the LR model displayed the best performance (AUC 0.950; F1 score 0.604; sensitivity 0.764; specificity 0.919) upon using the combination of clinical and imaging features.
Table 3. Results of using different feature sets (values in italics indicate the best results).

<table>
<thead>
<tr>
<th>Feature sets and model</th>
<th>Area under the curve</th>
<th>F1 score</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics + symptoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR^a</td>
<td>0.819</td>
<td>0.382</td>
<td>0.627</td>
<td>0.825</td>
</tr>
<tr>
<td>XGB^b</td>
<td>0.763</td>
<td>0.363</td>
<td>0.318</td>
<td>0.956</td>
</tr>
<tr>
<td>NN^c</td>
<td>0.730</td>
<td>0.332</td>
<td>0.427</td>
<td>0.880</td>
</tr>
<tr>
<td><strong>Clinical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>0.848</td>
<td>0.387</td>
<td>0.455</td>
<td>0.906</td>
</tr>
<tr>
<td>XGB</td>
<td>0.787</td>
<td>0.286</td>
<td>0.227</td>
<td>0.962</td>
</tr>
<tr>
<td>NN</td>
<td>0.647</td>
<td>0.237</td>
<td>0.309</td>
<td>0.881</td>
</tr>
<tr>
<td><strong>Imaging</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>0.926</td>
<td>0.593</td>
<td>0.818</td>
<td>0.901</td>
</tr>
<tr>
<td>XGB</td>
<td>0.904</td>
<td>0.486</td>
<td>0.636</td>
<td>0.896</td>
</tr>
<tr>
<td>NN</td>
<td>0.845</td>
<td>0.555</td>
<td>0.600</td>
<td>0.936</td>
</tr>
<tr>
<td><strong>Clinical + imaging</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>0.950</td>
<td>0.604</td>
<td>0.764</td>
<td>0.919</td>
</tr>
<tr>
<td>XGB</td>
<td>0.904</td>
<td>0.520</td>
<td>0.473</td>
<td>0.965</td>
</tr>
<tr>
<td>NN</td>
<td>0.782</td>
<td>0.413</td>
<td>0.486</td>
<td>0.907</td>
</tr>
</tbody>
</table>

^aLR: logistic regression.
^bXGB: XGBoost.
^cNN: neural network.

**Prediction at Baseline Severity With Oversampling**

Since the cohort was highly imbalanced, with the majority of cases being of mild/ordinary severity (imbalance ratio 0.07), we applied four oversampling methods to increase the ratio of severe/critical cases: SMOTE [38], Adaptive Synthetic sampling [39], geometric SMOTE [40], and CTGAN [41]. Figure 3 shows the differences in AUC values and F1 scores obtained through oversampling, with negative values indicating a reduction in AUC or F1 scores and positive values indicating the opposite trend. Oversampling resulted in greater improvements in the F1 score than in the AUC. The greatest improvement in the F1 score (0.09) was observed for the clinical features (clinical) with XGBoost and SMOTE (XGB-smo); however, the AUC decreased by 0.08 with the same method. Considering both AUC and F1 scores simultaneously, the combination of clinical and imaging features (clinical + imaging) benefited most from oversampling. In particular, the AUC and F1 score for clinical + imaging features were increased by 0.01 and 0.06, respectively, using LR with SMOTE (LR-smo).
Figure 3. Differences in the (A) area under the curve values and (B) F1 scores with oversampling and without oversampling. Positive values (blue) indicate oversampling resulting in higher values, negative values (red) indicating oversampling resulting in lower values. smo: synthetic minority oversampling; ada: Adaptive Synthetic sampling; geo = geometric synthetic minority oversampling; gan: conditional generative adversarial network; LR: logistic regression; NN: neural network; XGB: XGBoost.

Table 4 presents the best results of the evaluated models using various feature sets after oversampling. Oversampling did not improve the performance of the LR model for the demographic characteristics + symptoms features, but SMOTE and geometric SMOTE increased the F1 scores for clinical features and imaging features, respectively. Notably, the performance of the LR model (Table 3) was optimal for the combination of clinical and imaging features, with improvements in the AUC (0.960 vs 0.950), F1 score (0.668 vs 0.604), sensitivity (0.845 vs 0.764), and specificity (0.929 vs 0.919), after oversampling with SMOTE.

Table 4. The best results obtained using different feature sets after oversampling (arrow indicates improved performance after oversampling).

<table>
<thead>
<tr>
<th>Feature sets</th>
<th>Results</th>
<th>Area under the curve</th>
<th>F1</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics + symptoms</td>
<td>LR&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>0.819</td>
<td>0.382</td>
<td>0.627</td>
<td>0.825</td>
</tr>
<tr>
<td>Clinical</td>
<td>LR – smo&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.837</td>
<td>0.421</td>
<td>0.518&lt;sup&gt;↑&lt;/sup&gt;</td>
<td>0.902</td>
</tr>
<tr>
<td>Imaging</td>
<td>LR – geo&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.926</td>
<td>0.599</td>
<td>0.818</td>
<td>0.904&lt;sup&gt;↑&lt;/sup&gt;</td>
</tr>
<tr>
<td>Clinical + imaging</td>
<td>LR – smo</td>
<td>0.960&lt;sup&gt;↑&lt;/sup&gt;</td>
<td>0.668&lt;sup&gt;↑&lt;/sup&gt;</td>
<td>0.845&lt;sup&gt;↑&lt;/sup&gt;</td>
<td>0.929&lt;sup&gt;↑&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>LR: logistic regression.
<sup>b</sup>No improvement after oversampling.
<sup>c</sup>smo: synthetic minority oversampling.
<sup>d</sup>geo: geometric synthetic minority oversampling.

Model Interpretation

We used the SHAP framework [42] to interpret the output of the best-performing LR model through SMOTE oversampling. This framework helps determine the importance of a feature by comparing model predictions with or without the feature.

Figure 4 shows a SHAP plot summarizing how the values of each feature impact the model output of the LR model using all features (clinical and imaging features), with features sorted in descending order of importance. Figure 4A shows the feature importance scores sorted by the average impact on the model output, and Figure 4B shows the SHAP values of individual features. Furthermore, 4 imaging features, including consolidation volume (consolidation_val), total lesion volume (lesion_vol), ground-glass volume (groundglass_vol), and volume of other abnormalities (other_vol), are among the top 6 features, their high values increasing the likelihood of the model to predict a severe/critical COVID-19 case. Low albumin levels, high C-reactive protein levels, a high leukocyte count, and low lactate dehydrogenase levels make the model more likely to predict a critical/severe COVID-19 case. Moreover, older age and male gender increased the likelihood of the model to predict severe/critical COVID-19 cases.
Discussion

Principal Findings
In our cohort of patients with COVID-19, fever, cough, and fatigue were the most common symptoms, consistent with previous studies on COVID-19 [34]. The incidence of dyspnea and an increased respiratory rate was significantly higher in severe cases. Some symptoms such as sore muscle, headache, diarrhea, and nausea were present in 9–38 (2.6%-11.0%) of patients and did not differ significantly between mild and severe cases. Patients with severe COVID-19 tended to be of older age and had comorbidities (including cardiovascular disease, diabetes, hypertension, and cancer), concurrent with previous studies [1,3,5,34]. We observed no difference between males and females in our cohort, although the model did rely on gender for increasing the likelihood of predicting a severe/critical case.

A combination of clinical and imaging features yielded the best performance. Imaging features had the strongest impact on model output, with high values of consolidation volume, lesion volume, ground-glass volume, and other volume increasing the likelihood of the model to predict a severe case of COVID-19. Ground-glass opacity is an important feature of COVID-19 [14]. The inclusion of clinical features further improved the accuracy of severity assessment, with findings such as albumin levels, C-reactive protein levels, thromboplastin time, white blood cell counts, and lactate dehydrogenase levels being amongst the most informative features, concurrent with a previous study that also used laboratory findings to predict COVID-19–related mortality [9]. Furthermore, C-reactive protein was associated with a significant risk of critical illness in a study of 5279 patients with laboratory-confirmed COVID-19 [5]. Our model also relied on symptoms and patient characteristics such as gender, dyspnea, body temperature, diabetes, and respiratory rate for differentiating between mild and severe cases. Clinical features alone (demographics, signs, symptoms, and laboratory results), resulted in low sensitivity. Therefore, dependence on only clinical features poses the risk of predicting mild/ordinary COVID-19 among patients at the risk of critical/severe illness.

Oversampling yielded mixed results, although it revealed the best model performance in our study. The best model without oversampling (ie, the LR model) also yielded remarkable findings (AUC 0.950; F1 0.604; sensitivity 0.764; specificity 0.919), and SMOTE oversampling further improved the model performance (AUC 0.960; F1 0.668; sensitivity 0.845; specificity 0.929). Considering the propensity of health care data to be imbalanced [51-54], our results suggest the need for further analysis of oversampling methods for medical data sets. Self-supervision [55,56] may also help improve the performance of models using imbalanced medical data sets; in particular, future studies should evaluate the impact of self-supervision on tabular medical data.

Clinical Implications
The rapid spread of COVID-19 has overwhelmed health care systems, necessitating methods for efficient disease severity assessment. Our results indicate that clinical and imaging features can facilitate automated severity assessment of COVID-19. While our study would benefit from a larger data set, our results are encouraging because we trained the models with data from one hospital only and tested them using an independent data set from another hospital, albeit with high predictive accuracy.
The proposed methods and models would be useful in several clinical scenarios. First, the proposed models are fully automated and can expedite the assessment process, saving time in reading CT scans or evaluating patients through a scoring system. These models can be useful in hospitals that are overwhelmed by a high volume of patients during the outbreak by identifying severe cases as early as possible, such that treatment can be escalated. Our models, with low sensitivity and high specificity, are best used in combination with a model with high sensitivity and low specificity. A high-sensitivity model can identify patients with severe COVID-19, and our model (with high specificity) could identify false-positives; that is, patients with mild COVID-19 who were wrongly identified as having severe COVID-19.

Our models were developed and validated using 4 different feature sets, providing the flexibility to accommodate patients with different available data. For example, if a patient has neither a chest CT scan nor a blood test, the model based on demographics and symptoms can still achieve reasonably good prediction performance (AUC 0.819; sensitivity 0.627; specificity 0.825). Availability of the patients’ clinical and imaging features can improve the model’s sensitivity and specificity, with the potential to triage patients with COVID-19 (eg, prioritizing care for patients at a higher risk of mortality).

Limitations and Future Prospects

Our data set consisted of 346 patients with confirmed COVID-19, with data on 230 (66.5%) patients from HSCH used for training/validation and data on 116 (33.5%) patients from XYCH used for testing. Our data set was highly imbalanced, which could have made models overfit to the majority class. In addition, only the baseline data were used for patients in this study; therefore, we could not assess how early can COVID-19 progression be detected. We intend to further investigate the longitudinal data and design computational models to predict disease progression in our future studies.

While we explored various NN configurations, the results were not comparable to those of LR, presumably owing to the limited data set and the low dimensionality of the feature vectors. In this study, we used a complex NN model (EfficientNetB7 U-Net) to extract the imaging features and tested various models for classification using the combination of imaging features and tabular clinical data. Such 2-stage processing may simplify the classification task for these models, thereby reducing the need for another NN model for classification owing to low dimensionality of the features. Further exploration of NN architectures for tabular data would likely improve the performance of the NN model, especially if more data are available.

During training and validation, the performance of the models across cross-validation folds showed high variance owing to the small number of positive cases in the validation fold. A larger dataset would improve the reliability and robustness of the models. The data also consisted of COVID-19 cases which were confirmed through RT–PCR analysis of nasopharyngeal swabs. As such, our model is limited to differentiating severe/critical cases from mild/ordinary cases of COVID-19 and not for diagnosing COVID-19 or differentiating COVID-19 cases from those of other respiratory tract infections. Further studies are required to determine the efficacy of the severity assessments, including data from asymptomatic patients.

Using the Prediction Model Study Risk of Bias Assessment Tool [57], our models are at a high risk of bias owing to a potential bias in the participants domain (the cohort including participants [mean age 48.5 years, SD 15.4 years] who were admitted to hospitals) and the analysis domain (small sample size and class imbalance). Our models are at a low risk of bias in the predictor and outcome domains.

Conclusions

This study presents a novel method for severity assessment of patients diagnosed with COVID-19. Our results indicate that clinical and imaging features can be used for automated severity assessment of COVID-19. While imaging features had the strongest impact on the model’s performance, inclusion of clinical features and oversampling yielded the best performance in our study. The proposed method may potentially help triage patients with COVID-19 and prioritize care for patients at a higher risk of severe disease.

Acknowledgments

This study was supported by the Natural Science Foundation of Guangdong Province (grant number 2017A030313901), the Guangzhou Science, Technology and Innovation Commission (grant number 201804010239), the Foundation for Young Talents in Higher Education of Guangdong Province (grant number 2019KQNCX005), and the National Health and Medical Research Council (NHMRC) Centre of Research Excellence in Digital Health, and the NHMRC Partnership Centre for Health System Sustainability. SL acknowledges the support of an NHMRC grant (NHMRC Early Career Fellowship 1160760). We acknowledge Fujitsu Australia for providing the computational resources for this study.

Authors’ Contributions

SL, XQ, and X-RC share corresponding author responsibilities; SL (sidong.liu@mq.edu.au) will respond to technical inquiries, and XQ (qxm2020cov@163.com), and X-RC (caixran@jnu.edu.cn) will respond to inquiries regarding clinical data and applications.

Conflicts of Interest

None declared.

References

http://medinform.jmir.org/2021/2/e24572/


Abbreviations

- **AUC**: area under the curve
- **CT**: computed tomography
- **CTGAN**: conditional generative adversarial network
- **HSCH**: Huang Shi Central Hospital
- **LR**: logistic regression
- **NHMRC**: National Health and Medical Research Council
- **NN**: neural network
- **SHAP**: Shapley Additive Explanations
- **SMOTE**: synthetic minority oversampling
- **XYCH**: Xiang Yang Central Hospital

Please cite as:


Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study

JMIR Med Inform 2021;9(2):e24572
URL: http://medinform.jmir.org/2021/2/e24572/
doi: 10.2196/24572
PMID: