Original Paper

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 Quiroz[1,2*], PhD; You-Zhen Feng[3*], MD; Zhong-Yuan Cheng[3*], MD; Dana Rezazadegan[1,4], PhD; Ping-Kang Chen[3], MD; Qi-Ting Lin[3], MD; Long Qian[5], PhD; Xiao-Fang Liu[6,7], PhD; Shlomo Berkovsky[1], PhD; Enrico Coiera[1], PhD; Lei Song[8], MD; Xiaoming Qiu[9], MD; Sidong Liu[1], PhD; Xiang-Ran Cai[3], MD

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine, Health and Human Sciences, Macquarie University, Macquarie Park, Australia

[2]Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia

[3]Medical Imaging Centre, The First Affiliated Hospital of Jinan University, Guangzhou, China

[4]Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne, Australia

[5]Department of Biomedical Engineering, Peking University, Beijing, China

[6]Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Nankai University, Tianjin, China

[7]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

[8]Department of Radiology, Xiangyang Central Hospital, Affiliated Hospital of Hubei University of Arts and Science, Xiangyang, China

[9]Department 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).

XSL·FO
RenderX

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.

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.

XSL•FO

RenderX

Table 1. Demographic characteristics of the patients in the 2 cohorts (N=346).

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

[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.

Figure 1. Representative chest computed tomography scans and the lung and lesion models of (A) a mild COVID-19 case and (B) a severe COVID-19 case.

(a) Mild   (b) Severe
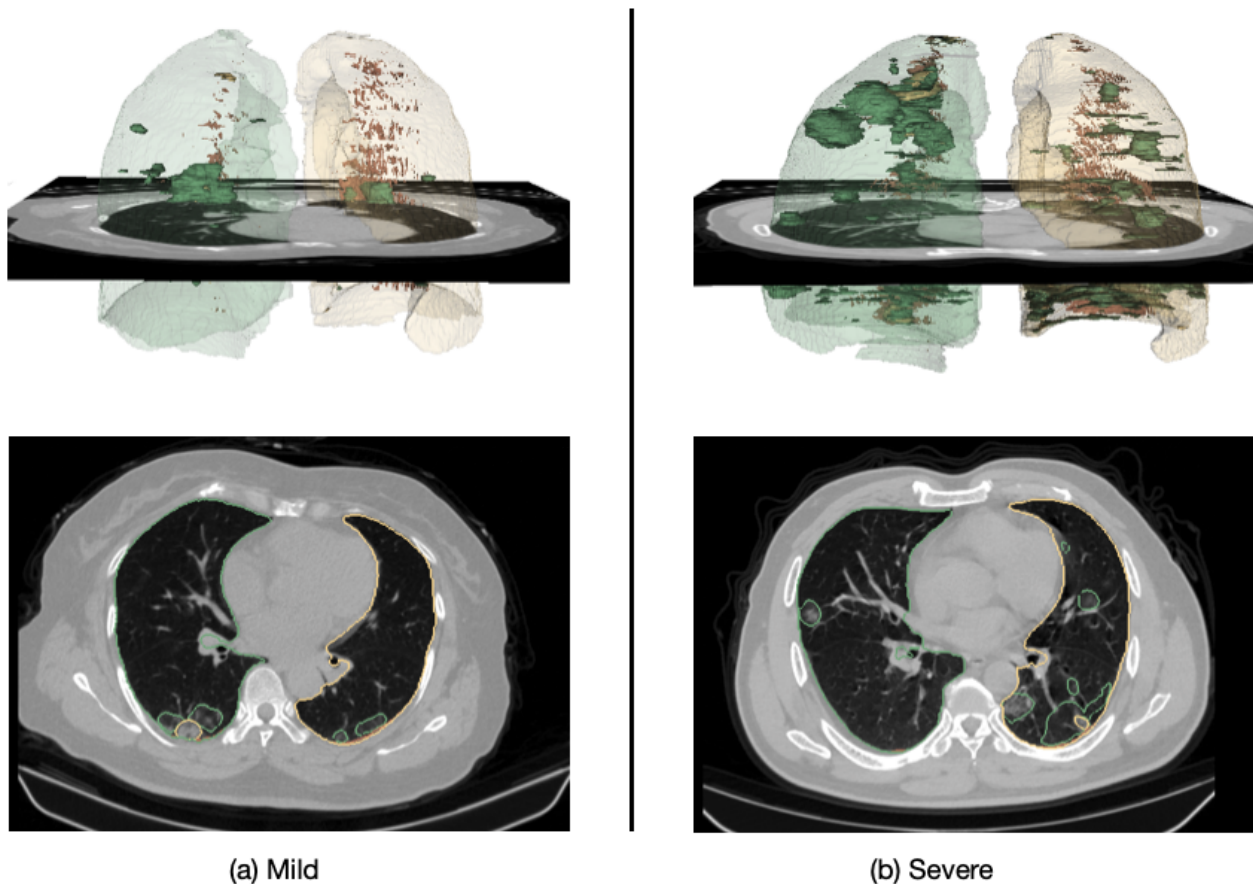
Clinical data collected from the patients included demographic characteristics, signs, symptoms, comorbidities, and the following 18 laboratory findings: white blood cell count ($\times 10^9$/L), neutrophil count ($\times 10^9$/L), lymphocyte count ($\times 10^9$/L), hemoglobin (g/L), platelets ($\times 10^9$/L), prothrombin time (s), activated partial thromboplastin time (s), D-dimer (nmol/L), C-reactive protein (mg/L), albumin (g/L), alanine aminotransferase (μkat/L), aspartate aminotransferase (μkat/L), total bilirubin (μmol/L), potassium (mmol/L), sodium (mmol/L), creatinine (μmol/L), creatine kinase (μkat/L), and lactate dehydrogenase (μkat/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.

Figure 2. Structure and dimensions of the feature sets. COPD: chronic obstructive pulmonary disorder.
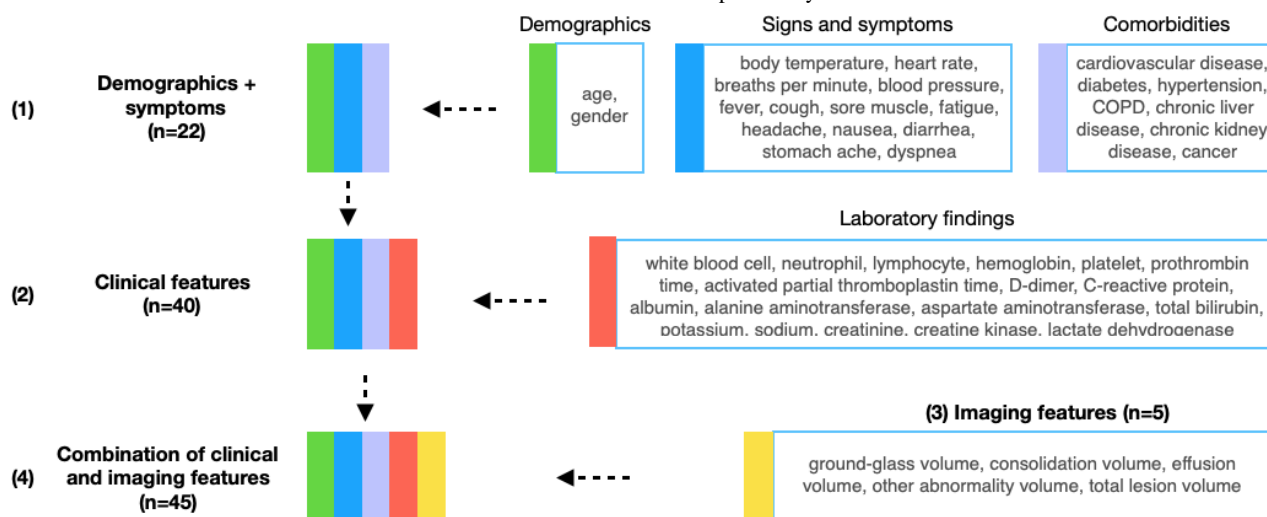
Severity Assessment Models

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 200×100 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.

Characteristics	Patients			P value[a]
	All patients	Mild/ordinary	Severe/critical	
Sample size, n	346	324	22	N/A[b]
Demographic characteristics				
Age (years), median (IQR)	49.0 (38.0-59.0)	48.5 (37.0-57.3)	63.0 (52.5-69.5)	<.001
Gender, n (%)				.38
Female	177 (51.2)	168 (51.9)	9 (41.0)	
Male	169 (48.8)	156 (48.1)	13 (59.0)	
Comorbidities, n (%)				
Cardiovascular disease	40 (11.6)	32 (9.9)	8 (36.0)	.002
Diabetes	34 (9.8)	28 (8.6)	6 (27.0)	.01
Hypertension	51 (14.7)	42 (13.0)	9 (41.0)	.002
Chronic obstructive pulmonary disease	11 (3.2)	9 (2.8)	2 (9.0)	.15
Chronic liver disease	7 (2.0)	7 (2.2)	0 (0)	N/A
Chronic kidney disease	4 (1.2)	3 (0.9)	1 (5.0)	.20
Cancer	8 (2.3)	5 (1.5)	3 (14.0)	.01
Signs, median (IQR)				
Body temperature	37.8 (37-38.3)	37.8 (37-38.3)	38.1 (37.1-39)	.11
Heart rate	85 (80-90)	85 (80-90)	90 (80-101.8)	.11
Breaths per minute	20 (20-21)	20 (20-21)	21 (20-28)	.002
Blood pressure high	120 (119.5-130.0)	120 (118.5-130.0)	127 (120-146.5)	.07
Blood pressure low	74 (69-80)	74 (69-80)	79.5 (71-89)	.08
Symptoms, n (%)				
Fever	275 (79.5)	256 (79.0)	19 (86.0)	.59
Cough	238 (68.8)	220 (67.9)	18 (82.0)	.24
Fatigue	118 (34.1)	108 (33.3)	10 (45.0)	.25
Dyspnea	32 (9.2)	23 (7.1)	9 (41.0)	<.001
Sore muscle	38 (11.0)	35 (10.8)	3 (14.0)	.72
Headache	34 (9.9)	31 (9.6)	3 (14.0)	.47
Diarrhea	23 (6.6)	20 (6.2)	3 (14.0)	.17
Nausea	9 (2.6)	7 (2.2)	2 (9.0)	.11

[a]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.

[b]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).

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

[a]LR: logistic regression.

[b]XGB: XGBoost.

[c]NN: 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).

XSL·FO

RenderX

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.
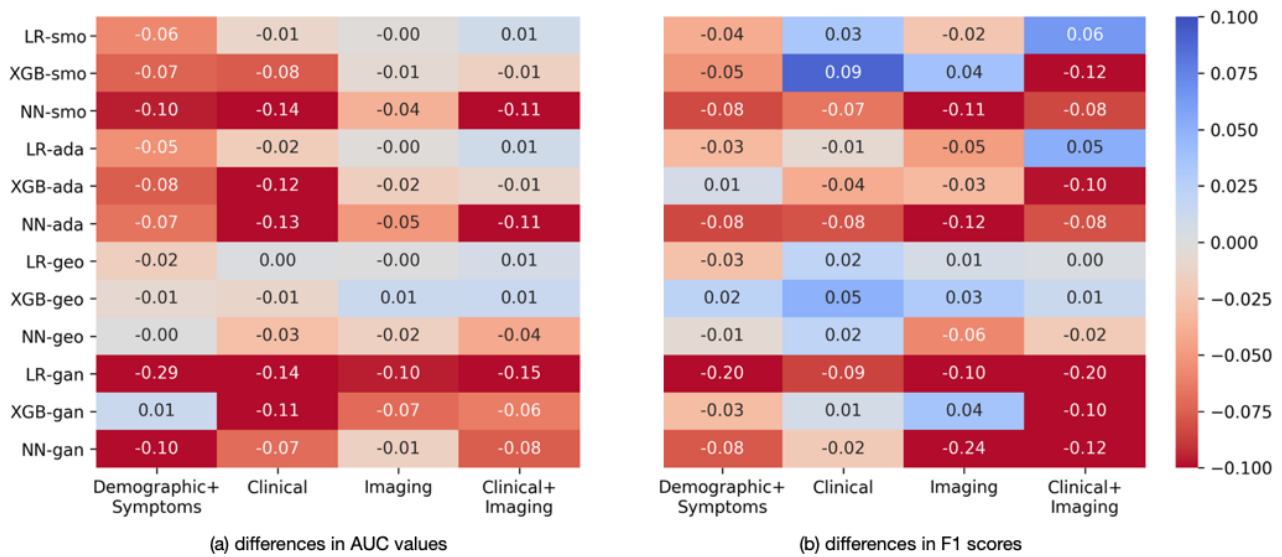

(a) differences in AUC values

(b) differences in F1 scores

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).

Feature sets	Results				
	Model	Area under the curve	F1	Sensitivity	Specificity
Demographics + symptoms	LR[a,b]	0.819	0.382	0.627	0.825
Clinical	LR – smo[c]	0.837	0.421 ↑	0.518 ↑	0.902
Imaging	LR – geo[d]	0.926	0.599 ↑	0.818	0.904 ↑
Clinical + imaging	LR – smo	0.960 ↑	0.668 ↑	0.845 ↑	0.929 ↑

[a]LR: logistic regression.

[b]No improvement after oversampling.

[c]smo: synthetic minority oversampling.

[d]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.

Figure 4. (A) Feature importance, evaluated using the mean SHAP (Shapely Addictive Explanations) values, in the logistic regression (LR) model using all features. (B) SHAP plot for the LR model using all features. Each point represents a feature instance, and the color indicates the feature value (red: high, blue: low). Negative SHA*P* values indicate feature instances contributing to a model output of a mild/ordinary COVID-19 case, whereas positive SHA*P* values indicate features contributing to a model output of a severe/critical COVID-19 case.
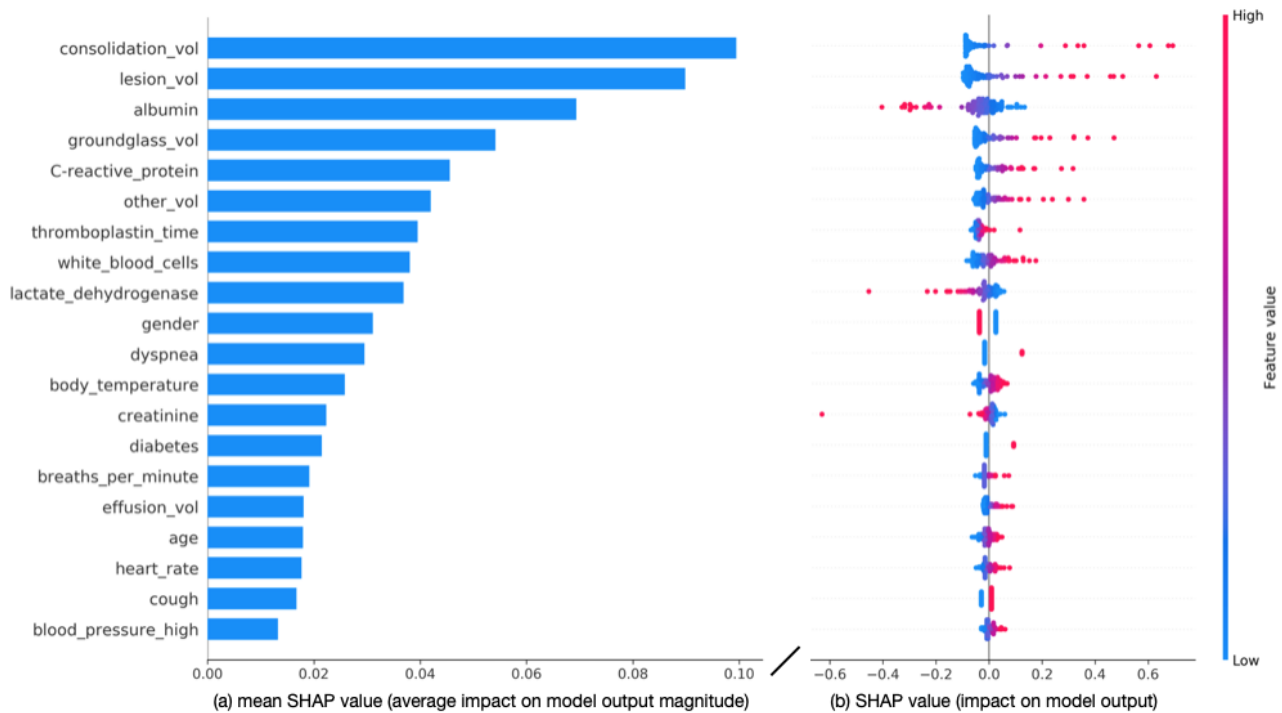
Discussion

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 for patients were used 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.

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

1. Siordia J. Epidemiology and clinical features of COVID-19: A review of current literature. J Clin Virol 2020 Jun;127:104357 [FREE Full text] [doi: 10.1016/j.jcv.2020.104357] [Medline: 32305884]

2. Angulo FJ, Finelli L, Swerdlow DL. Reopening Society and the Need for Real-Time Assessment of COVID-19 at the Community Level. JAMA 2020 Jun 09;323(22):2247-2248. [doi: 10.1001/jama.2020.7872] [Medline: 32412582]

3. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, the Northwell COVID-19 Research Consortium, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA 2020 May 26;323(20):2052-2059 [FREE Full text] [doi: 10.1001/jama.2020.6775] [Medline: 32320003]

4. Madjid M, Safavi-Naeini P, Solomon SD, Vardeny O. Potential Effects of Coronaviruses on the Cardiovascular System: A Review. JAMA Cardiol 2020 Jul 01;5(7):831-840. [doi: 10.1001/jamacardio.2020.1286] [Medline: 32219363]

5. Petrilli CM, Jones SA, Yang J, Rajagopalan H, O'Donnell L, Chernyak Y, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. BMJ 2020 May 22;369:m1966 [FREE Full text] [doi: 10.1136/bmj.m1966] [Medline: 32444366]

6. Jin X, Pang B, Zhang J, Liu Q, Yang Z, Feng J, et al. Core Outcome Set for Clinical Trials on Coronavirus Disease 2019 (COS-COVID). Engineering (Beijing) 2020 Oct;6(10):1147-1152 [FREE Full text] [doi: 10.1016/j.eng.2020.03.002] [Medline: 32292626]

7. Inui S, Fujikawa A, Jitsu M, Kunishima N, Watanabe S, Suzuki Y, et al. Chest CT Findings in Cases from the Cruise Ship with Coronavirus Disease (COVID-19). Radiology: Cardiothoracic Imaging 2020 Apr 01;2(2):e200110-e201180 [FREE Full text] [doi: 10.1148/ryct.2020200110] [Medline: 32565152]

8. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology 2020 Feb 26:200642. [doi: 10.1148/radiol.2020200642] [Medline: 32101510]

9. Yan L, Zhang H, Xiao Y, Wang M, Sun C, Liang J, et al. A Machine Learning-Based Model for Survival Prediction in Patients With Severe COVID-19 Infection. medRxiv Preprint posted online March 17, 2020. [FREE Full text] [doi: 10.1101/2020.02.27.20028027]

10. Butt C, Gill J, Chun D, Babu BA. Deep learning system to screen coronavirus disease 2019 pneumonia [retracted in: Appl Intell. 2020]. Appl Intell 2020 Apr 22. [doi: 10.1007/s10489-020-01714-3]

11. Chen J, Wu L, Zhang J, Zhang L, Gong D, Zhao Y, et al. Deep Learning-Based Model for Detecting 2019 Novel Coronavirus Pneumonia on High-Resolution Computed Tomography: A Prospective Study. medRxiv Preprint posted online March 01, 2020. [FREE Full text] [doi: 10.1101/2020.02.25.20021568]

12. Gozes O, Frid-Adar M, Greenspan H, Browning P, Zhang H, Ji W, et al. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection and Patient Monitoring using Deep Learning CT Image Analysis. arXiv Preprint posted online March 24, 2020 [FREE Full text]

13. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology 2020 Aug;296(2):E65-E71 [FREE Full text] [doi: 10.1148/radiol.2020200905] [Medline: 32191588]

14. Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, et al. Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images. medRxiv Preprint posted online February 25, 2020. [FREE Full text] [doi: 10.1101/2020.02.23.20026930]

15. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. A Deep Learning Algorithm Using CT Images to Screen for Coronavirus Disease (COVID-19). medRxiv Preprint posted online April 24, 2020. [doi: 10.1101/2020.02.14.20023028]

16. Feng Y, Liu S, Cheng Z, Quiroz J, Rezazadegan D, Chen P, et al. Severity Assessment Progression Prediction of COVID-19 Patients Based on the LesionEncoder FrameworkChest CT. medRxiv Preprint posted online August 06, 2020. [doi: 10.1101/2020.08.03.20167007]

17. Chaganti S, Balachandran A, Chabin G, Cohen S, Flohr T, Georgescu B, et al. Quantification of Tomographic Patterns associated with COVID-19 from Chest CT. arXiv Preprint posted online April 2, 2020. [Medline: 32550252]

18. Wong A, Lin Z, Wang L, Chung A, Shen B, Abbasi A, et al. Towards computer-aided severity assessment: training and validation of deep neural networks for geographic extent and opacity extent scoring of chest X-rays for SARS-CoV-2 lung disease severity. arXiv Preprint posted online August 22, 2020. [FREE Full text]

19. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. Radiology 2020 Sep;296(3):E156-E165 [FREE Full text] [doi: 10.1148/radiol.2020201491] [Medline: 32339081]

20. Ozturk T, Talo M, Yildirim E, Baloglu U, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020 Jun;121:103792 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103792] [Medline: 32568675]

21. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 2020 Nov 11;10(1):19549 [FREE Full text] [doi: 10.1038/s41598-020-76550-z] [Medline: 33177550]

22. Mei X, Lee H, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020 Aug;26(8):1224-1228 [FREE Full text] [doi: 10.1038/s41591-020-0931-3] [Medline: 32427924]

23. Xiao L, Li P, Sun F, Zhang Y, Xu C, Zhu H, et al. Development and Validation of a Deep Learning-Based Model Using Computed Tomography Imaging for Predicting Disease Severity of Coronavirus Disease 2019. Front Bioeng Biotechnol 2020;8:898 [FREE Full text] [doi: 10.3389/fbioe.2020.00898] [Medline: 32850746]

24. Qian X, Fu H, Shi W, Chen T, Fu Y, Shan F, et al. M Lung-Sys: A Deep Learning System for Multi-Class Lung Pneumonia Screening From CT Imaging. IEEE J Biomed Health Inform 2020 Dec;24(12):3539-3550. [doi: 10.1109/JBHI.2020.3030853] [Medline: 33048773]

25. Hu S, Gao Y, Niu Z, Jiang Y, Li L, Xiao X, et al. Weakly Supervised Deep Learning for COVID-19 Infection Detection and Classification From CT Images. IEEE Access 2020 Aug;8(8):118869-118883. [doi: 10.1109/ACCESS.2020.3005510] [Medline: 33156775]

26. Wang Z, Liu Q, Dou Q. Contrastive Cross-Site Learning With Redesigned Net for COVID-19 CT Classification. IEEE J Biomed Health Inform 2020 Oct;24(10):2806-2813. [doi: 10.1109/JBHI.2020.3023246] [Medline: 32915751]

27. Babukarthik RG, Adiga VAK, Sambasivam G, Chandramohan D, Amudhavel J. Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN). IEEE Access 2020;8:177647-177666. [doi: 10.1109/access.2020.3025164]

28. Ng M, Lee EYP, Yang J, Yang F, Li X, Wang H, et al. Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review. Radiology: Cardiothoracic Imaging 2020 Feb 01;2(1):e200034. [doi: 10.1148/ryct.2020200034]

29. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]

30. Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, et al. Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. IEEE Rev Biomed Eng 2021;14:4-15. [doi: 10.1109/RBME.2020.2987975] [Medline: 32305937]

31. Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z, Hadjilooei F, Lalbakhsh P, et al. Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment. IEEE Access 2020;8:109581-109595. [doi: 10.1109/ACCESS.2020.3001973]

32. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020 Oct;139:110059 [FREE Full text] [doi: 10.1016/j.chaos.2020.110059] [Medline: 32834612]

33. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [doi: 10.1016/S0140-6736(20)30183-5] [Medline: 31986264]

34. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis 2020 Apr;20(4):425-434 [FREE Full text] [doi: 10.1016/S1473-3099(20)30086-4] [Medline: 32105637]

35. Bishop C. In: Jordan M, Kleinberg J, Schölkopf B, editors. Pattern Recognition and Machine Learning (Information Science and Statistics). New York, NY: Springer-Verlag; 2006.

36. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. New York, NY: Association for Computing Machinery; 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2016; New York, NY. [doi: 10.1145/2939672.2939785]

37. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull Math Biol 1990;52(1-2):99-115; discussion 73. [Medline: 2185863]

38. Chawla NV, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]

39. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 Presented at: IEEE International Joint Conference on Neural Networks; 1-8 June 2008; Hong Kong URL: https://ieeexplore.ieee.org/document/4633969 [doi: 10.1109/IJCNN.2008.4633969]

40. Douzas G, Bacao F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. Information Sciences 2019 Oct;501:118-135 [FREE Full text] [doi: 10.1016/j.ins.2019.06.007]

41. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. arXiv Preprint posted online October 28, 2019. [FREE Full text]

42. Lundberg S, Lee S. A Unified Approach to Interpreting Model Predictions. arXiv Preprint posted online November 25, 2017. [FREE Full text]

43. Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, et al. Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. Sci Rep 2020 May 07;10(1):7733 [FREE Full text] [doi: 10.1038/s41598-020-64588-y] [Medline: 32382048]

44. Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 2020 Aug 20;4(1):50 [FREE Full text] [doi: 10.1186/s41747-020-00173-2] [Medline: 32814998]

45. COVID-19 CT segmentation dataset. MedSeg. 2020. URL: http://medicalsegmentation.com/covid19/ [accessed 2021-02-05]

46. Kikinis R, Pieper S, Vosburgh K. 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support. In: Intraoperative Imaging and Image-Guided Therapy. New York, NY: Springer; 2013:277-289.

47. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825-2830.

48. Howard J, Gugger S. Fastai: A Layered API for Deep Learning. Information 2020 Feb 16;11(2):108 [FREE Full text] [doi: 10.3390/info11020108]

49. Smith L. Cyclical Learning Rates for Training Neural Networks. : IEEE; 2017 Presented at: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV); 24-31 March 2017; Santa Rosa, CA. [doi: 10.1109/wacv.2017.58]

50. Youden WJ. Index for rating diagnostic tests. Cancer 1950 Jan;3(1):32-35. [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]

51. Krawczyk B, Schaefer G, Woźniak M. A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification. Artif Intell Med 2015 Nov;65(3):219-227. [doi: 10.1016/j.artmed.2015.07.005] [Medline: 26319694]

52. Jiang J, Liu X, Zhang K, Long E, Wang L, Li W, et al. Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. Biomed Eng Online 2017 Nov 21;16(1):132 [FREE Full text] [doi: 10.1186/s12938-017-0420-1] [Medline: 29157240]

53. Gan D, Shen J, An B, Xu M, Liu N. Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Computers & Industrial Engineering 2020 Feb;140:106266 [FREE Full text] [doi: 10.1016/j.cie.2019.106266]

54. Zhang L, Yang H, Jiang Z. Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. Biomed Eng Online 2018 Dec 04;17(1):181 [FREE Full text] [doi: 10.1186/s12938-018-0604-3] [Medline: 30514298]

55. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013 Aug;35(8):1798-1828. [doi: 10.1109/TPAMI.2013.50] [Medline: 23787338]

56. Raina R, Battle A, Lee H, Packer B, Ng A. Self-Taught Learning: Transfer Learning From Unlabeled Data. New York, NY: Association for Computing Machinery; 2007 Presented at: ICML '07: Proceedings of the 24th International Conference on Machine learning; June 2007; Corvalis, OR. [doi: 10.1145/1273496.1273592]

57. Wolff R, Moons K, Riley R, Whiting P, Westwood M, Collins G, PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019 Jan 01;170(1):51-58 [FREE Full text] [doi: 10.7326/M18-1376] [Medline: 30596875]

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

XSL•FO
RenderX

XSL•FO

RenderX