Severity Assessment of COVID-19 based on Clinical and Imaging Data

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

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

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

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

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

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

[6] Department of Radiology, Huangshi Central Hospital, Affiliated Hospital of Hubei Polytechnic University, Edong Healthcare Group, Huangshi, China

[7] Centre for Big Data Research in Health, UNSW, Sydney, Australia

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

* Equal contribution

Corresponding author

ABSTRACT

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

Materials and Methods Clinical data—demographics, signs, symptoms, comorbidities and blood test results—and chest CT scans of 346 patients from two hospitals in the Hubei province, China, were used to develop machine learning models for automated severity assessment of diagnosed COVID-19 cases. We compared the predictive power of clinical and imaging data by testing multiple machine learning models, and further explored the use of four oversampling methods to address the imbalance distribution issue. Features with the highest predictive power were identified using the SHAP framework.

Results Targeting differentiation between mild and severe cases, logistic regression models achieved the best performance on clinical features (AUC:0.848, sensitivity:0.455, specificity:0.906), imaging features (AUC:0.926, sensitivity:0.818, specificity:0.901) and the combined features (AUC:0.950, sensitivity:0.764, specificity:0.919). The SMOTE oversampling method further improved the performance of the combined features to AUC of 0.960 (sensitivity:0.845, specificity:0.929).

Discussion 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 findings from previous studies. Oversampling yielded mixed results, although it achieved the best performance in our study.

Conclusions This study indicates that clinical and imaging features can be used for automated severity assessment of COVID-19 patients and have the potential to assist with triaging COVID-19 patients and prioritizing care for patients at higher risk of severe cases.

KEYWORDS

COVID-19, severity assessment, clinical features, CT scans, imbalanced data, oversampling

BACKGROUND AND SIGNIFICANCE

Coronavirus disease 2019 (COVID-19) has overwhelmed health systems worldwide.[1,2] As of July 5, 2020, more than 11 million cases have been confirmed worldwide, with 528,953 global deaths.[3] Given the various complications associated with COVID-19,[4-6] methods that facilitate triage of COVID-19 can help prioritize care for those who are likely to experience severe or critical cases. COVID-19 illness severity can be defined as four groups: (1) mild, (2) ordinary, (3) severe, and (4) critical.[7] Severe and critical cases require intensive care and more healthcare resources. A high rate of false positive severe or critical cases could overwhelm healthcare resources (i.e., ICU beds). Equally important, delays in identifying severe or critical cases would lead to delayed treatment of patients at a higher risk of mortality. It is, therefore, important to identify severe cases as early as possible, so that resources can be mobilized and treatment can be escalated.

Chest CT scans have been found to provide important diagnostic and prognostic information,[8,9] and consequently, they have been the focus of numerous recent studies using machine learning techniques for prediction tasks related to the COVID-19 pandemic.[10-22] Studies have looked at mortality predictions,[10] diagnosis (detecting COVID cases and differentiating from other pulmonary diseases or no disease),[11,15–19] and severity assessment and disease progression.[20–22] The majority of current approaches have used deep learning and imaging features from CT-scans[11,12,15–19] and X-rays,[13,14,22] with popular architectures including ResNet,[11,16,18] U-Net,[15,21] Inception,[19] Darknet,[13] and other convolutional neural networks.[14,22] More details can be found in recent review papers[1,23–25].

Automatic assessment of chest CT scans to predict COVID-19 severity is of a great clinical importance, but has only been the focus of few studies.[20–22] Automated assessment of chest CT scans can substantially reduce the image reading time for radiologists, provide quantitative

information that can be compared across patients and time-points, with clinical applications in detection and diagnosis, progression tracking and prognosis.[9] While CT scans are an important diagnostic tool, prior work has also shown that clinical data, such as symptoms, comorbidities and laboratory findings, differed for COVID-19 patients who were admitted to intensive care units (ICU) vs non-ICU patients,[26] and were predictive of the mortality risk.[10] One study compared the imaging data and clinical data of 81 confirmed COVID-19 patients, suggesting that the combination of imaging features with clinical and laboratory findings could facilitate early diagnosis of COVID-19.[27]

In this study, we used patient clinical data and imaging data to predict COVID-19 case severity. We consider this as a binary classification task, predicting whether a diagnosed patient is likely to develop a mild or a severe case of COVID. The contributions of this work are three-fold. First, we compared the predictive power of clinical and image data for severity assessment, by testing three machine learning models: logistic regression (LR),[28] gradient boosted trees (XGB),[29] and neural network (NN).[30] Secondly, due to the cohort data being highly imbalanced, with the majority of cases being mild/ordinary, we explored the use of four oversampling methods to address the imbalance distribution issue.[31–34] Third, we interpreted the importance of features using the SHAP (SHapley Additive exPlanations) framework and identified the features with the highest predictive power.[35] The evaluated predictive models yielded high accuracy and identified predictive imaging and clinical features consistent with prior findings.

MATERIALS AND METHODS

Study Design and Participants

This is a retrospective study carried out with data collected by two hospitals in the Hubei province, China. The study cohort consisted of patients who had COVID-19 diagnosis

confirmed by reverse transcription PCR (RT-PCR). A total of 346 patients from two hospitals were retrospectively enrolled, including 230 patients from Huang Shi Central Hospital (HSCH) and 116 from Xiang Yang Central Hospital (XYCH). These patients were admitted to hospital between 11-01-2020 and 23-02-2020, and all underwent chest CT scans at initial hospitalization. All the participants provided written consent. This study was approved by the Institutional Review Board of both hospitals. **Table 1** shows the demographics of the two cohorts of patients.

Table 1. Demographics of the two cohorts of patients.

Category	HSCH	XYCH	Total
Mild	7	1	8
Ordinary	212	104	316
Severe	7	6	13
Critical	4	5	9
Total patients	230	116	346
Age (mean±SD)	49.0±14.4	47.5±17.2	48.5±15.4
Gender (female/male)	120/110	57/59	177/169

Imaging and Clinical Data

Chest CT scans were collected from the patients at initial hospitalization. All CT scans were pre-processed with intensity normalization, contrast limited adaptive histogram equalization, and gamma adjustment, using the same pre-processing pipeline as in our previous study.[36] We performed lung segmentation on the CT slices using an established model - R231CovidWeb,[37] trained on a large and diverse dataset of non-COVID-19 CT scans and further fine-tuned with an additional COVID-19 dataset.[38] The CT slices with less than $3mm^2$ lung tissue were removed from the datasets since they bear little or no information of the lung. The lesions were segmented using EfficientNetB7 U-Net,[20] also trained using a public COVID-19 dataset.[38] The model produced four types of lesions, including 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, which were further normalized by the lung volume. **Figure 1** shows examples of the lung and lesion segmentation results of a mild case and a severe case. The upper row presents the 3D models of the lung and lesions, reconstructed using 3D Slicer (v4.6.2),[39] and the lower row presents the axial CT slices with the lung and lesion (green: ground-glass opacities; yellow: consolidation; brown: pleural effusion) boundaries overlaid on the CT slices.

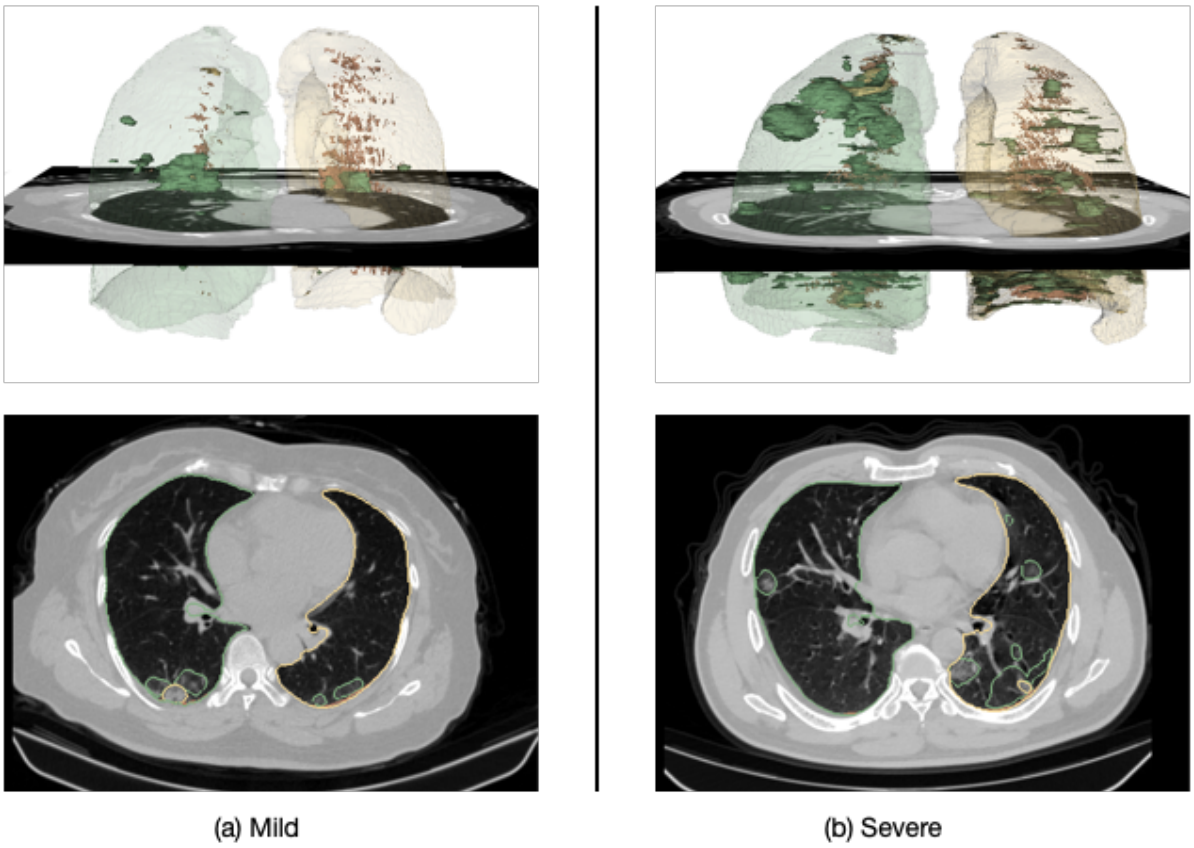

Figure 1. Examples of the CT scans and the lung and lesion models of (a) mild case and (b) a severe case.

Clinical data collected from the patients included demographics, signs, symptoms, comorbidities and 18 laboratory test results: white blood cell count ($\times 10^9$/L), neutrophil count

($\times 10^9$/L), lymphocyte count ($\times 10^9$/L), haemoglobin ($\times 10^9$/L), platelet ($\times 10^9$/L), prothrombin time (s), activated partial thromboplastin time (s), D-dimer (mg/L), C-reactive protein (mg/L), albumin (g/L), alanine aminotransferase (U/L), aspartate aminotransferase (U/L), total bilirubin (mmol/L), potassium (mmol/L), sodium (mmol/L), creatinine (μmol/L), creatine kinase (U/L), and lactate dehydrogenase (U/L).

All the features were either continuous or binary—all binary features belong to signs, symptoms and comorbidities. Continuous features were standardized to be centred around 0 with a standard deviation of 1. **Figure 2** shows the structure and dimensions of the features used in this study.

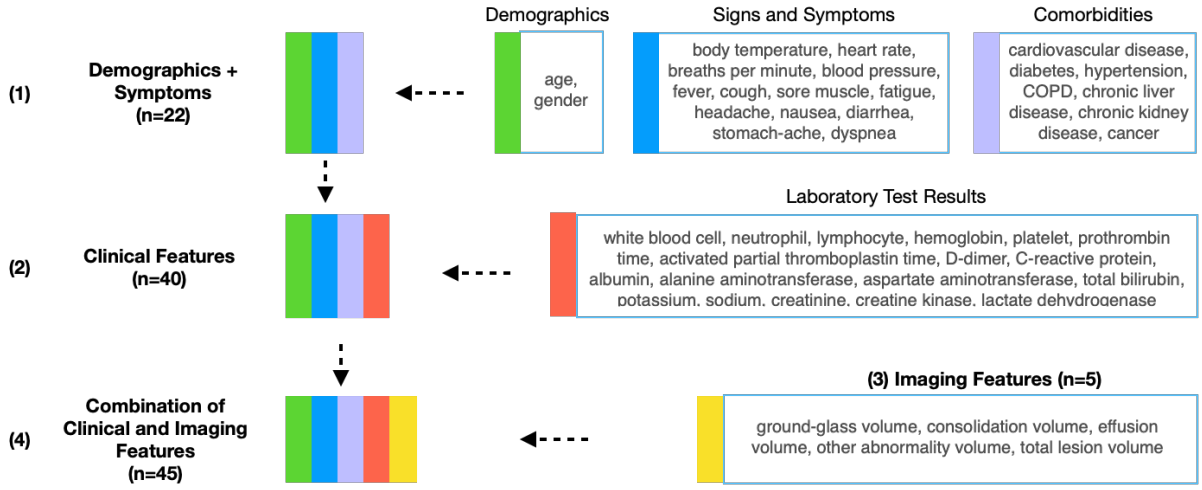

Figure 2. Structure and dimensions of the feature sets.

These features were grouped into four feature sets: (1) demographic and symptoms (a subset of the available clinical features), (2) clinical features (demographic, signs and symptoms, and laboratory test results), (3) imaging features extracted from the CT chest scans using deep learning, and (4) combination of clinical and imaging features.

Severity Assessment Models

Three models were trained and compared to predict case severity: logistic regression (with Scikit-learn),[40] gradient boosted trees (with XGBoost),[29] and a neural network (with

Fastai).[41] We used the HSCH data (230 samples) for training and validation using 5-fold repeated stratified cross-validation. The XYCH data (116 samples) was withheld for testing. We report results for the test set using AUC and F1 scores averaged over the independent runs.

Hyperparameter exploration and tuning were done using the train/validation set. Random search was used to tune the hyperparameters of LR and XGB. For NN, we used a four-layer, fully-connected architecture, with the first hidden layer having 200 nodes and the second hidden layer having 100 nodes. The learning rate (0.01) was determined using the learning rate finder.[42] All other parameters of NN were set to default values. We explored a different number of nodes in the first and second hidden layers, with 200x100 yielding the best results in the validation set. Out of the 346 patients, 167 (48%) had at least one missing feature (5.7 on average, mostly in the laboratory test results category). Missing feature values were imputed with the mean for each feature.

Oversampling

The majority of cases in our dataset were mild/ordinary cases and the minority were severe/critical cases. The imbalance ratio for the entire dataset was 0.07, for the training/validation set – 0.05, and for the testing set – 0.10. We tested four oversampling methods to increase the ratio of the minority class: Synthetic Minority over-Sampling Technique (SMOTE),[31] ADAptive SYNthetic sampling approach (ADASYN),[32] geometric SMOTE,[33] and a conditional generative adversarial network model for tabular data (CTGAN).[34] For all these methods, we oversampled the training set, trained a model on the oversampled data, and reported results on the same test set. We fixed the resampling ratio of all methods to 0.3 (bringing imbalance ratio to 0.3). When using CTGAN for oversampling, we fitted the CTGAN model with the training set, sampled to generate synthetic data, keeping only synthetic data for the minority class (severe/critical), repeating until the ratio of the minority to the majority class reached 0.3.

RESULTS

Patient Characteristics

Table 2 presents the patients' characteristics. The differences between the mild/ordinary and severe/critical groups were calculated with the Mann-Whitney U test and Fisher's exact test. In the full cohort, the median age was 49 (IQR 38-59). The median age for patients with mild/ordinary cases was 48.5 (IQR 37 – 57.3) and for the severe/critical cases it was 63 (IQR 52.5 – 69.5). There are statistically significant differences between patients with severe/critical and mild/normal infections with respect to age ($P < 0.001$) and comorbidities of cardiovascular disease ($P=0.002$), hypertension ($P=0.002$), diabetes ($P=0.01$), and cancer ($P=0.01$). Out of all the signs and symptoms, raised respiration rate ($P=0.002$) and dyspnea ($P<0.001$) were more common in patients with severe/critical cases of COVID-19.

Table 2. Demographics and baseline characteristics of patients with confirmed COVID-19. Patients with higher severity of COVID-19 were more likely to have cardiovascular disease and experience shortness of breath as a symptom.

Characteristics		All patients (n=346)	Mild/Ordinary (n=324)	Severe/Critical (n=22)	p-value
Demographics					
Age		49 (38 – 59)	48.5 (37 – 57.3)	63 (52.5 – 69.5)	< 0.001
Gender	Female	177 (51%)	168 (52%)	9 (41%)	0.38
	Male	169 (49%)	156 (48%)	13 (59%)	
Comorbidities					
Cardiovascular disease		40 (12%)	32 (10%)	8 (36%)	0.002
Diabetes		34 (10%)	28 (9%)	6 (27%)	0.01
Hypertension		51 (15%)	42 (13%)	9 (41%)	0.002
COPD		11 (3%)	9 (3%)	2 (9%)	0.152
Chronic liver disease		7 (2%)	7 (2%)	0 (0%)	-
Chronic kidney disease		4 (1%)	3 (1%)	1 (5%)	0.20

Cancer	8 (2%)	5 (2%)	3 (14%)	0.01
Signs				
Body temperature	37.8 (37– 38.3)	37.8 (37 – 38.3)	38.1 (37.1 – 39)	0.11
Heart rate	85 (80 – 90)	85 (80 – 90)	90 (80 – 101.8)	0.11
Breaths per minute	20 (20 – 21)	20 (20 – 21)	21 (20 – 28)	0.002
Blood pressure high	120 (119.5–130)	120 (118.5–130)	127 (120–146.5)	0.07
Blood pressure low	74 (69 – 80)	74 (69 – 80)	79.5 (71 – 89)	0.08
Symptoms				
Fever	275 (79%)	256 (79%)	19 (86%)	0.59
Cough	238 (69%)	220 (68%)	18 (82%)	0.24
Fatigue	118 (34%)	108 (33%)	10 (45%)	0.25
Dyspnea	32 (9%)	23 (7%)	9 (41%)	< 0.001
Sore muscle	38 (11%)	35 (11%)	3 (14%)	0.72
Headache	34 (10%)	31 (10%)	3 (14%)	0.47
Diarrhea	23 (7%)	20 (6%)	3 (14%)	0.17
Nausea	9 (3%)	7 (2%)	2 (9%)	0.11
Stomach-ache	0 (0%)	0 (0%)	0 (0%)	-

*Continuous variables are expressed as median with lower and upper quartiles. Binary variables are expressed as n (%). P-values comparing mild/ordinary and severe/critical cases were obtained with Mann-Whitney U test and Fisher's exact test. As no patient in our cohort had stomach-ache, this feature was not used in our modelling.

Prediction of COVID-19 Severity at Baseline

The data from HSCH (230 patients) was used for training and validation, and the data from XYCH (116 patients) was used as the independent test set. We compared model performance using four feature sets: (1) demographics and symptoms, (2) clinical features, (3) imaging features and (4) combination of clinical and imaging features, as shown in **Figure 2**. The optimal classification threshold for the sensitivity, specificity and F1 score was identified using Youden's index.[43] **Table 3** shows the severity assessment performance of an LR model, an XGB 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 LR only. Hence, LR yielded the best performance (AUC = 0.950, F1 Score = 0.604, sensitivity = 0.764, specificity = 0.919) using the combination of clinical and imaging features.

Table 3. Results from using different feature sets.

	Model	AUC	F1	Sensitivity	Specificity
Demographic + Symptoms	LR	**0.819**	**0.382**	**0.627**	0.825
	XGB	0.763	0.363	0.318	**0.956**
	NN	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

LR = logistic regression; XGB = gradient boosted trees; NN = neural network. Bold-faced values indicate the best results.

Prediction at Baseline Severity with Oversampling

Since the cohort was highly imbalanced, with the majority of cases being mild/ordinary (imbalance ratio of 0.07), we applied four oversampling methods to increase the ratio of severe/critical cases: SMOTE,[31] ADASYN,[32] geometric SMOTE,[33] and CTGAN.[34] **Figure 3** shows the differences in AUC values and F1 scores resulting from the use of oversampling, with negative values indicating a decrease in AUC or F1 score and positive values indicating the opposite. Oversampling resulted in greater improvements in F1 score

compared to AUC. The greatest improvement in F1 (0.09) is observed for the clinical features (Clinical) with XGB and SMOTE method (XGB-smo); however, the AUC dropped by 0.08 with the same method. Considering both AUC and F1 score at the same time, the combination of clinical and imaging features (Clinical + Imaging) benefited the most from oversampling. Specifically, the AUC and F1 score for Clinical + Imaging features were increased by 0.01 and 0.06, respectively, using LR with SMOTE (LR-smo).

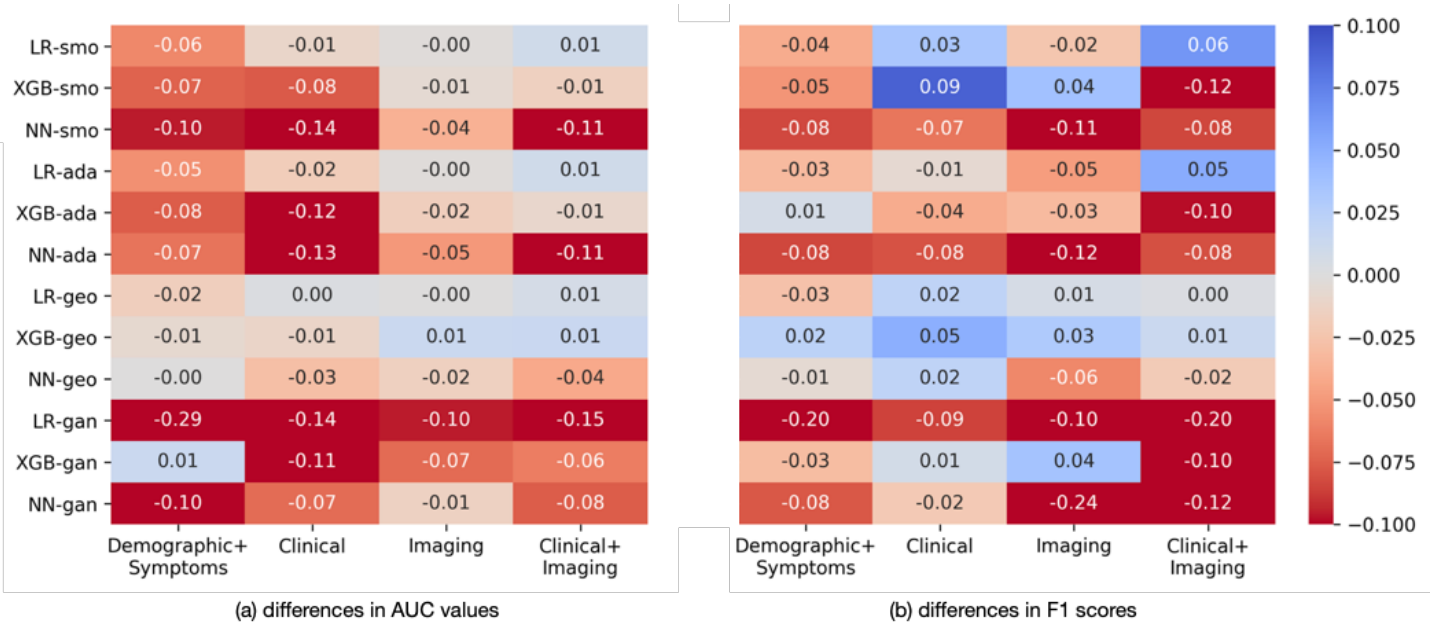

Figure 3. Differences in (a) AUC values and (b) F1 scores with oversampling compared to those without oversampling. Positive values (blue) indicate oversampling resulting in higher values, negative values (red) indicate oversampling resulting in lower values. smo = SMOTE; ada = ADASYN; geo = geometric SMOTE; gan = CTGAN; LR = logistic regression, NN = neural network, XGB = gradient boosted trees.

Table 4 presents the best results of the evaluated models using various feature sets after oversampling. Oversampling did not improve LR's performance on Demographics + Symptoms features, but SMOTE and geometric SMOTE resulted in increased F1 scores using Clinical features and Imaging features, respectively. Notably, the best performing in **Table 3** LR model using a combination of clinical and imaging features further improved to AUC of

0.960 (vs. 0.950), F1 score of 0.668 (vs. 0.604), sensitivity of 0.845 (vs. 0.764) and specificity of 0.929 (vs. 0.919), after oversampling with SMOTE.

Table 4. The best results from using different feature sets after oversampling.

	Model	AUC	F1	Sensitivity	Specificity
Demographic + Symptoms	LR*	0.819	0.382	0.627	0.825
Clinical	LR – smo	0.837	0.421 ↑	0.518 ↑	0.902
Imaging	LR – geo	0.926	0.599 ↑	0.818	0.904 ↑
Clinical + Imaging	LR – smo	0.960 ↑	0.668 ↑	0.845 ↑	0.929 ↑

smo = SMOTE; geo = geometric SMOTE; LR = logistic regression; *no improvement after oversampling; ↑improved performance after oversampling.

Model Interpretation

We used the SHAP (SHapley Additive exPlanations) framework[35] to interpret the output of the best performing LR model with SMOTE oversampling. This framework calculates the importance of a feature by comparing model predictions with and without the feature. **Figure 4** illustrates 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 from most important to least important. **Figure 4(a)** shows feature importance scores sorted by the average impact on the model output, and **Figure 4(b)** presents the SHAP values of individual feature instances. Four 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 six features with their high values resulting in the model being more likely to predict a severe/critical case of COVID-19. Low albumin, high counts of C-reactive protein, high counts of leukocytes, and low values of lactate dehydrogenase make the model more likely to predict a case severity of critical/severe. Older age and male gender also made the model more likely to predict severe/critical cases.

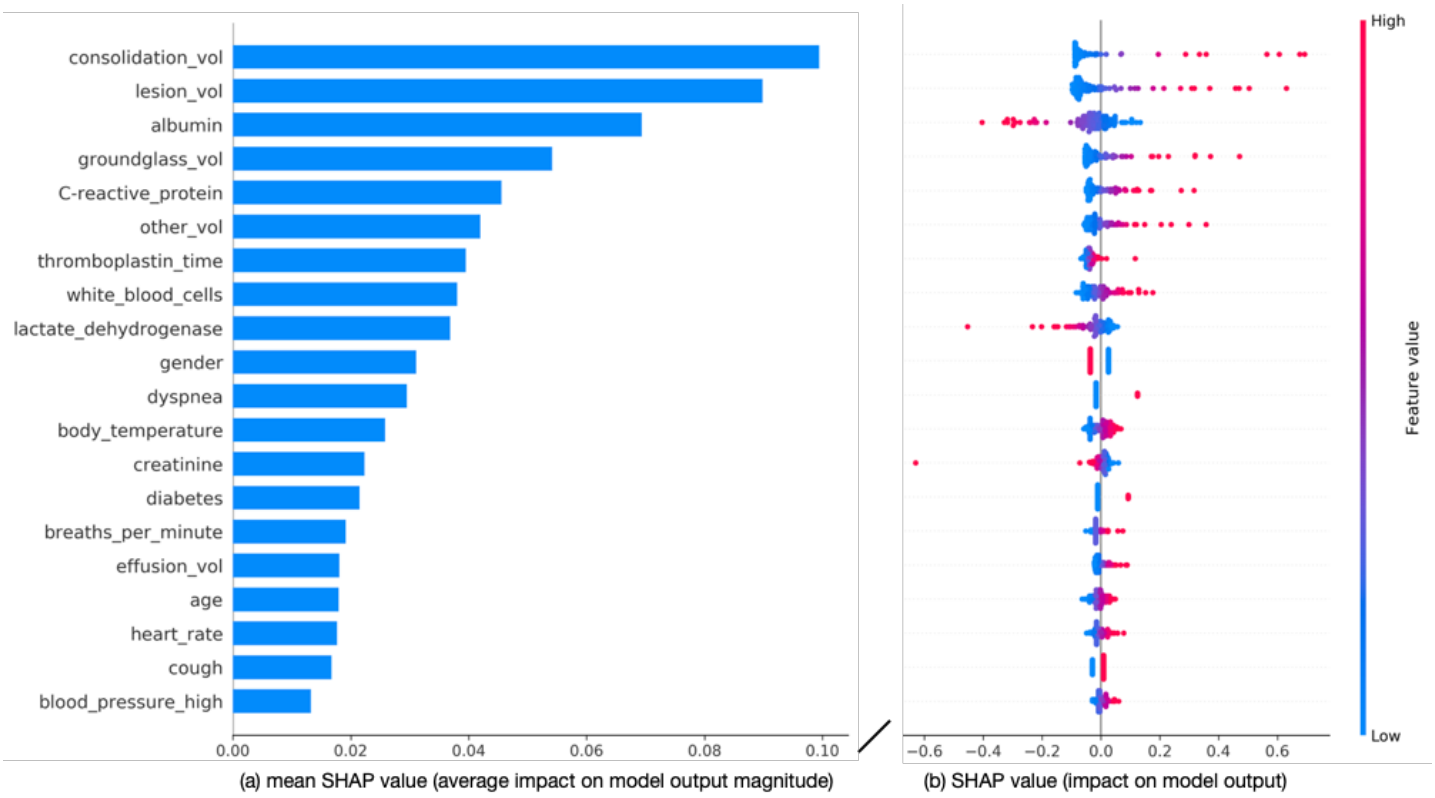

Figure 4. (a) Feature importance, evaluated using the mean SHAP values, in the LR model using all features. (b) SHAP plot for the LR model using all features. Each point represents a feature instance, and the color indicates the feature value (red for high and blue for low). Negative SHAP values indicate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

DISCUSSION

Main Findings

In our cohort, fever, cough, and fatigue were the most common symptoms in patients with COVID-19, consistent with other studies of COVID-19 patients.[27] Severe cases manifested a statistically higher incidence of dyspnea and raised respiratory rate. Some symptoms such as sore muscle, headache, diarrhea, and nausea were present in 3-11% of patients and were not statistically different between mild and severe cases. Severe cases of COVID-19 tended to be of older age and had medical comorbidities (cardiovascular disease, diabetes, hypertension, cancer), in similar to prior studies.[1,4,6,27] There was 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 the model output, with high values of consolidation volume, lesion volume, ground-glass volume, and other volume making the model more likely to predict a severe case of COVID-19. Ground-glass opacity has also been found to be an important feature in prior work.[18] Inclusion of clinical features further improved the accuracy of severity assessments, with findings such as albumin, C-reactive protein, thromboplastin time, white blood cell counts, and lactate dehydrogenase being amongst the most informative features. The identification of lactate dehydrogenase, white blood cell counts, and C-reactive protein as informative features is consistent with findings from one prior study that also used laboratory findings for COVID-19 mortality prediction.[10] C-reactive protein was also associated with a significant risk of critical illness in a study of 5,279 laboratory confirmed COVID-19 patients.[6] Symptoms and patient characteristics such as gender, dyspnea, body temperature, diabetes, and breaths per minute were also relied on by the model for differentiating between mild and severe cases. Clinical features alone (demographics, signs, symptoms, and laboratory results), resulted in low sensitivity. Relying only on clinical features, thus, poses the risk of predicting mild/ordinary severity for patients who will develop a critical/severe case of COVID-19.

Oversampling yielded mixed results, although it resulted in the best model performance in our study. We note that the best model without oversampling (LR) also yielded strong results (AUC: 0.950, F1: 0.604, sensitivity: 0.764, specificity: 0.919), and the SMOTE oversampling method improved the performance further (AUC: 0.960, F1: 0.668, sensitivity: 0.845, specificity: 0.929). Given the propensity of imbalanced data in healthcare,[44–47] our results suggest the need for further analysis of oversampling methods for medical datasets. Self-

supervision,[48,49] may also help in improving performance on imbalanced medical datasets; in particular, future work should evaluate the impact of self-supervision on tabular medical data.

Clinical Implications

The rapid spread of COVID-19 has put a strain on healthcare systems, necessitating efficient disease severity assessment of COVID-19 patients. Results from this study indicate that clinical and imaging features can inform automated severity assessment of COVID-19. While our work would benefit from a larger dataset, our current results are encouraging given that the models were trained on data from one hospital only and tested on an independent dataset from another hospital, demonstrating nevertheless strong 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 the CT scans or evaluating patients using a scoring system. They can be of use in hospitals that are overwhelmed by a high volume of patients during the outbreak by identifying severe cases as early as possible, so that treatment can be escalated. Our models, with a higher specificity and relatively lower sensitivity, would best be used in combination with a model with higher sensitivity in diagnostic situations, i.e., a high sensitivity model can identify the patients with severe cases of COVID-19, and our model (with high specificity) could reduce false positives—patients with a mild case of COVID-19 who were wrongly identified as having a severe case of COVID-19.

Our models were developed and validated on four different feature sets, providing the flexibility to accommodate patients with different available data. For example, if a patient does not have 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). If the clinical and imaging features are available for patients, the model's sensitivity and specificity can be improved, with potential in triaging of COVID-19 patients, e.g., prioritizing care for patients at a higher risk of mortality.

Limitations

Our dataset consisted of 346 patients with confirmed COVID-19, with the data of 230 patients from the HSCH hospital used for training/validation and the 116 patients from the XYCH hospital used for testing. Our dataset 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 the progression can be detected. We will be further investigating the longitudinal data and designing computational models to predict disease progression in our future work.

While we explored various configurations of NN, results were not comparable to LR, presumably due to the limited dataset 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 imaging features combined with tabular clinical data. Such two-stage process may simplify the classification task for these models, thereby reducing the need for another NN model for classification due to low dimensionality of features. Further exploration of NN architectures for tabular data is likely to benefit the performance of the NN model, especially if more data is available.

During training and validation, the performance of the models across cross-validation folds showed high variance due 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 with RT-PCR. 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 for differentiating COVID-19 cases from other respiratory tract infections. Further work is needed to determine the efficacy of the severity assessments, including data from asymptomatic patients.

CONCLUSIONS

This work presents a novel method for severity assessment of diagnosed COVID-19 patients. The results indicate that clinical and imaging features can be used for automated severity assessment of COVID-19 patients. 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 have the potential to assist with triaging COVID-19 patients and prioritizing care for patients at higher risk of severe cases.

REFERENCES

[1] J. A. Siordia, "Epidemiology and clinical features of COVID-19: A review of current literature," *J. Clin. Virol.*, vol. 127, p. 104357, 2020.

[2] F. J. Angulo, L. Finelli, and D. L. Swerdlow, "Reopening Society and the Need for Real-Time Assessment of COVID-19 at the Community Level," *JAMA*, vol. 323, no. 22, pp. 2247–2248, Jun. 2020.

[3] "Johns Hopkins Coronavirus Resource Center." [Online]. Available: https://coronavirus.jhu.edu/map.html. [Accessed: 30-Jun-2020].

[4] S. Richardson *et al.*, "Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area," *JAMA*, vol. 323, no. 20, pp. 2052–2059, May 2020.

[5] M. Madjid, P. Safavi-Naeini, S. D. Solomon, and O. Vardeny, "Potential Effects of Coronaviruses on the Cardiovascular System: A Review," *JAMA Cardiol.*, Mar. 2020.

[6] C. M. Petrilli *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*, vol. 369, 2020.

[7] X. Jin *et al.*, "Core Outcome Set for Clinical Trials on Coronavirus Disease 2019 (COS-COVID)," *Engineering*, 2020.

[8] S. Inui *et al.*, "Chest CT Findings in Cases from the Cruise Ship 'Diamond Princess' with Coronavirus Disease 2019 (COVID-19)," *Radiol. Cardiothorac. Imaging*, vol. 2, no. 2, p. e200110, Mar. 2020.

[9] T. Ai *et al.*, "Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," *Radiology*, p. 200642, Feb. 2020.

[10] L. Yan *et al.*, "Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan," *medRxiv*, p. 2020.02.27.20028027, Jan. 2020.

[11] C. Butt, J. Gill, D. Chun, and B. A. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia," *Appl. Intell.*, pp. 1–7, Apr. 2020.

[12] H. X. Bai *et al.*, "AI Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Etiology on Chest CT," *Radiology*, p. 201491, Apr. 2020.

[13] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, p. 103792, 2020.

[14] L. Wang and A. Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images." 2020.

[15] J. Chen *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, p. 2020.02.25.20021568, Jan. 2020.

[16] O. Gozes *et al.*, "Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis." 2020.

[17] L. Li *et al.*, "Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT," *Radiology*, p. 200905, Mar. 2020.

[18] Y. Song *et al.*, "Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images," *medRxiv*, p. 2020.02.23.20026930, Jan. 2020.

[19] S. Wang *et al.*, "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," *medRxiv*, p. 2020.02.14.20023028, Jan. 2020.

[20] Y.-Z. Feng *et al.*, "Severity Assessment and Progression Prediction of COVID-19 Patients based on the LesionEncoder Framework and Chest CT," 2020.

[21] S. Chaganti *et al.*, "Quantification of Tomographic Patterns associated with COVID-19 from Chest CT." 2020.

[22] A. Wong *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." 2020.

[23] M.-Y. Ng *et al.*, "Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review," *Radiol. Cardiothorac. Imaging*, vol. 2, no. 1, p. e200034, Feb. 2020.

[24] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, 2020.

[25] F. Shi *et al.*, "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, pp. 1–1, 2020.

[26] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, Feb. 2020.

[27] H. Shi *et al.*, "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study," *Lancet Infect. Dis.*, vol. 20, no. 4, pp. 425–434, Apr. 2020.

[28] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[29] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[30] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.

[31] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority over-Sampling Technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[32] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.

[33] G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Inf. Sci. (Ny).*, vol. 501, pp. 118–135, 2019.

[34] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular data using Conditional GAN," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 7335–7345.

[35] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[36] S. Liu *et al.*, "Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning," *Sci. Rep.*, vol. 10, no. 1, p. 7733, 2020.

[37] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem." 2020.

[38] "COVID-19 CT segmentation dataset." [Online]. Available: http://medicalsegmentation.com/covid19/.

[39] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, "3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support BT - Intraoperative Imaging and Image-Guided Therapy," F. A. Jolesz, Ed. New York, NY: Springer New York, 2014, pp. 277–289.

[40] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in {P}ython," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[41] J. Howard and S. Gugger, "Fastai: A Layered API for Deep Learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[42] L. N. Smith, "Cyclical Learning Rates for Training Neural Networks." 2015.

[43] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, Jan. 1950.

[44] B. Krawczyk, G. Schaefer, and M. Woźniak, "A hybrid cost-sensitive ensemble for imbalanced breast thermogram classification," *Artif. Intell. Med.*, vol. 65, no. 3, pp. 219–227, 2015.

[45] J. Jiang *et al.*, "Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network," *Biomed. Eng. Online*, vol. 16, no. 1, p. 132, 2017.

[46] D. Gan, J. Shen, B. An, M. Xu, and N. Liu, "Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis," *Comput. Ind. Eng.*, vol. 140, p. 106266, 2020.

[47] L. Zhang, H. Yang, and Z. Jiang, "Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN," *Biomed. Eng. Online*, vol. 17, no. 1, p. 181, 2018.

[48] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[49] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-Taught Learning: Transfer Learning from Unlabeled Data," in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 759–766.

ACKNOWLEDGMENTS

This project was supported by Natural Science Foundation of Guangdong Province (grant no.2017A030313901); Guangzhou Science, Technology and Innovation Commission (grant no.201804010239); Foundation for Young Talents in Higher Education of Guangdong Province (grant no.2019KQNCX005); and the NHMRC Centre of Research Excellence in Digital Health and the NHMRC Partnership Centre for Health System Sustainability. Dr Sidong Liu acknowledges the support of an Australian National Health and Medical Research Council grant (NHMRC Early Career Fellowship 1160760). We acknowledge Fujitsu Australia Limited for providing the computational resources for this study.

COMPETING INTERESTS STATEMENT

The authors have no conflict of interest nor any competing interests to declare.

CONTRIBUTORSHIP STATEMENT

The project was initially conceptualized and supervised by X.R.C., S.L, X.M.Q. and L.S. The patient data and imaging data were acquired by Y.Z.F., Z.Y.C. and D.R. The analysis methods were designed and implemented by J.C.Q., S.L. and H.X. The data were analyzed by J.C.Q. and S.L. The research findings were interpreted by P.K.C., Q.T.L, L.Q., X.F.L, S.B. and E.C. All authors were involved in the design of the work. The manuscript was drafted by J.C.Q., S.L and L.Q., and all authors have substantively revised it. All authors have reviewed and approved the submitted version.