# A Quantitative Lung Computed Tomography Image Feature for Multi-Center Severity Assessment of COVID-19

Biswajoy Ghosh[1,*], Nikhil Kumar[1], Anup K. Sadhu[2], Nirmalya Ghosh[1], Pabitra Mitra[1], and Jyotirmoy Chatterjee[1]

[1]Indian Institute of Technology Kharagpur, India
[2]EKO CT & MRI Scan Centre, Medical College and Hospitals Campus, Kolkata, India
[*]correspondence email: biswajoyghosh@smst.iitkgp.ac.in

June 13, 2020

**Abstract**

The COVID-19 pandemic has affected millions and congested healthcare systems globally. Hence an objective severity assessment is crucial in making therapeutic decisions judiciously. Computed Tomography (CT)-scans can provide demarcating features to identify severity of pneumonia —commonly associated with COVID-19—in the affected lungs. Here, a quantitative severity assessing chest CT image feature is demonstrated for COVID-19 patients. An open-source multi-center Italian database[1] was used, among which 60 cases were incorporated in the study (age 27-86, 71% males) from 27 CT imaging centers. Lesions in the form of opacifications, crazy-paving patterns, and consolidations were segmented. The severity determining feature —$L_{norm}$ was quantified and established to be statistically distinct for the three —mild, moderate, and severe classes (p-value<0.0001). The thresholds of $L_{norm}$ for a 3-class classification were determined based on the optimum sensitivity/specificity combination from Receiver Operating Characteristic (ROC) analyses. The feature $L_{norm}$ classified the cases in the three severity categories with 86.88% accuracy. 'Substantial' to 'almost-perfect' intra-rater and inter-rater agreements were achieved involving expert and non-expert based evaluations ($\kappa$-score 0.79-0.97). We trained machine learning based classification models and showed $L_{norm}$ alone has a superior diagnostic accuracy over standard image intensity and texture features. Classification accuracy was further increased when $L_{norm}$ was used for 2-class classification i.e. to delineate the severe cases from non-severe ones with a high sensitivity (97.7%), and specificity (97.49%). Therefore, key highlights of this severity assessment feature are accuracy, lower dependency on expert availability, and wide utility across different imaging centers.

*Keywords:* COVID-19, Computed Tomography, severity assessment, lung, machine learning; health informatics

## 1 Introduction

With the onset of the COVID-19 pandemic caused by the SARS-CoV-2 coronavirus, newer tools and techniques are increasingly needed for efficient detection and therapy. As of now, RT-PCR based detection of the virus from oral/nasal swabs is globally accepted as the confirmatory test. However, due to the chances of the absence of viral particles on the swab especially in the asymptomatic or mild cases, the sensitivity of the method suffers (71% [1]). Therefore several screening methods are being deployed to augment COVID-19 detection including clinical history, symptom assessment, blood tests and imaging methods. Among the different imaging methods, chest X-Ray [2], Computed Tomography (CT) [3], and Ultrasonography (USG) [4, 5] are used in different clinical settings across the world to identify features associated with lung pneumonia commonly caused in the COVID-19 infection.

[1]`https://www.sirm.org/category/senza-categoria/covid-19/`

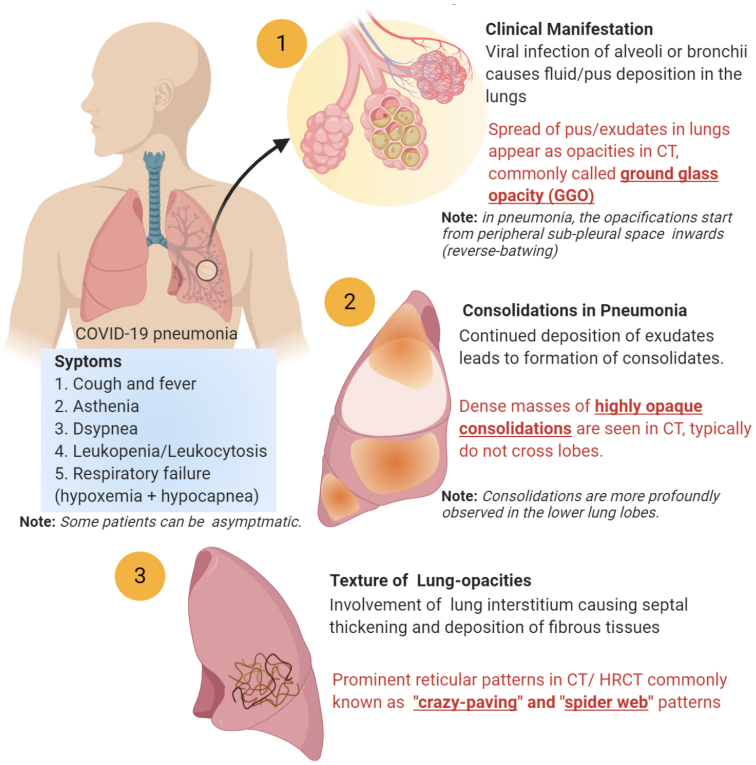


Figure 1: Illustration of pathological features of lung pneumonia in COVID-19 patients along with corresponding CT features.

Among these imaging methods, CT and high-resolution CT (HRCT) have shown a sensitivity of up to 98% [1] and hence has emerged as a strong screening tool for COVID-19 [3]. Thus, CT-imaging reduces the incidence of several infected individuals being discharged back in the community [6, 7].

The most common pathology seen in COVID-19 is pneumonia [7] (Figure 1) which eventually disrupts the lungs' ability for gaseous exchange and reduces the oxygen availability for the normal cells to function. Due to the excessive deposition of fluid and pus (exudates) in the alveolar space, breathing is obstructed leading to respiratory failure. In severe cases, the immune response goes systemic and damages other vital organs like heart [8], liver [9] and kidney [10]. This increases the mortality of people with already present underlying health conditions and comorbidities like CAD (coronary artery disease), COPD (chronic obstructive pulmonary disease), CRF (chronic renal failure), diabetes, cancer, hepatitis, immunodeficiency etc. Although the disease can be fatal, most individuals show only mild symptoms and do not need hospitalization.

Due to the entire range of symptoms expressing in the population from asymptomatic to fatal, severity assessment is crucial for effective administration of the right therapeutic drugs as per the patient's condition [11]. This becomes even more complicated with underlying conditions as some drugs that are otherwise effective for the virus, may have adverse effects on pre-existing conditions. Currently severity assessment is done by symptoms and chemical tests (liver function, pO2, saO2, procalcitonin, troponin, creatinine, blood cell count, inflammatory markers etc.). However most of the specific markers are expressed late into the disease and do not provide the direct status of the most affected organ, i.e. the lungs.

Some recent literature have shown chest CT to determine severity. Yang et. al. have analysed 20 segments from chest CT images (with constant CT parameters) manually (by expert) and based on an objective rating of opacification cover in each segment have assessed mild and severe disease state [12]. Although the evaluation is clinically thorough, the method is expertise-heavy and the requirement for a manual scoring of 20 different segments to obtain a severity score makes prediction of an incoming case cumbersome. Shen et. al. on the other hand have used (1) lesion percentage cover and (2) mean lesion density to evaluate severity in three categories in 44 affected individuals [13]. They have used a computed aided tool to semi-automatically segment lesions and identify the correlation between

the two parameters and chest CT pathological features. Although, the computer aided performance correlated well with expert performance, no severity scoring index or predictive performance for classification into severity groups have been shown. Huang et. al. have used U-Net deep learning architecture in segmentation of the lung lesions [14], measured the percent opacification cover, and have shown that this feature was statistically different for the four severity classes. The segmentation method has a high performance and is a major highlight in automated assessment of opacification cover. However, the paper does not address the predictive performance of the feature in classifying severity, nor provides clear feature thresholds in differentiating the severity groups.

In this paper, we demonstrate chest CT image feature and associated framework for a 3-class severity assessment in COVID-19 positive individuals.

### 1.1 Logical Exposition

Recent literature have correlated chest CT image features with the corresponding pathological disease severity [13, 15]. These gradual emergence of pneumonia attributes —ground-glass opacifications, reticular patterns, and dense consolidations [16] in CT images were found to correlate with exudate accumulation and septal thickening/lung fibrosis affecting breathing. Since these pathologies causes a higher absorbance of X-rays (and hence shows a higher CT value) [17], the disease severity should correspond strongly to the lesion gray scale intensity. However, the gray scale-intensity has been found only to be moderately correlated to disease severity [13]. This is because CT scans are often modified among CT centers. Also, individual CT scanners can be set to variably operate at specific CT window settings (level and width) as per user's requirement. Additionally, the X-ray tube settings, rotation time, pitch, slice thickness etc. are some of the other variables that can affect the perception of lesion 'density' in a CT slice and hence severity analyses. Also, image contrast-enhancement for better visualization post-imaging is often performed affecting the CT-values further. Thus a type of normalization of the lesion gray scale intensity is required i.e. independent of the amount of post-processing and parametric variables adopted in CT imaging. Although use of multiple features for severity assessment is important, but multi-variate analysis, use of complex and black-box methods of classification often affect the interpretation of the findings especially for a disease like COVID-19 where images are used to indirectly interpret the pathological severity. Here we have shown an anatomically normalized intensity as a lesion feature and relevant framework to assess severity of lung pneumonia in COVID-19 patients. Comparative evaluation with other methods and features, methodological validations, and relevant thresholds have been also demonstrated.

## 2 Methods

A schematic methodological work-flow has been illustrated in Fig. 2.

### 2.1 Dataset

All the CT images used for the method has been taken from the repository of the Italian Society of Medical and Interventional Radiology (SIRM) [18]. CT and HRCT images of 60 COVID-19 positive individuals were taken between the age of 27 to 86 (39 males, 15 females, gender not specified for 6 cases)(Supplementary Table S1). No pediatric or COVID-19 negative CT images were included in the study. All clinical annotations given with the cases were used as well. The database includes Multi-center CT scans from 27 different CT/ HRCT imaging parameters (Supplementary Table S2).

### 2.2 Ground Truth for COVID-19 Severity

All the lung-CT images were evaluated by a radiology expert given the case history, CT images, and corresponding radiological findings in the case-reports to assign three classes based on severity. The severity was divided into three classes —mild (S-1), moderate (S-2), and severe (S-3) [19] (Figure 3). It is to be noted that S-3 class includes severe and above (including critical) cases.

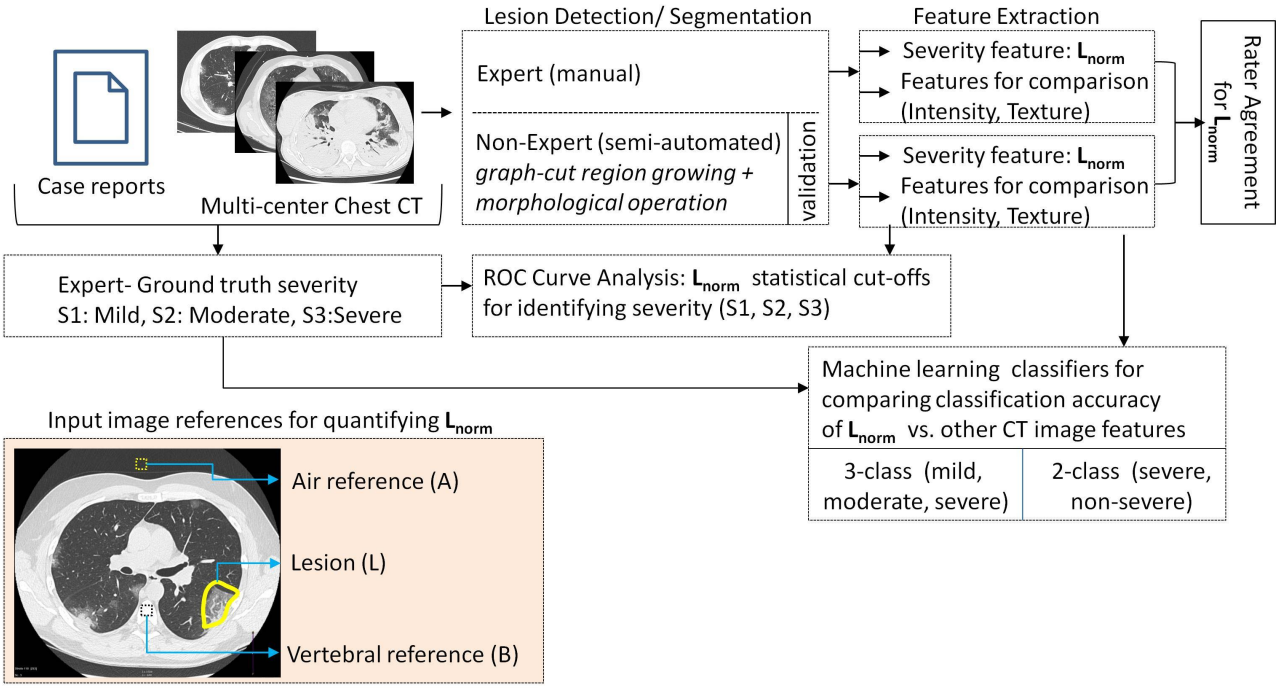


Figure 2: Schematic flow of $L_{norm}$ feature quantification for severity analysis of COVID-19 affected lungs from multi-center chest CT data. The scheme illustrates the framework for the application and validation of $L_{norm}$.

## 2.3 Lesion Detection

### 2.3.1 Expert (manual lesion area annotation)

A radiological expert was provided chest CT images of COVID-19 patients and was asked to manually draw the boundaries of the lung lesions. A total of 60 images were manually annotated by the expert and was used for feature extraction.

### 2.3.2 Non-expert (semi-automatic lesion detection)

a. All images were converted to 8-bit gray scale format. No image pre-processing was performed.
b. Lesion detection- Graph-cut based adaptive region growing algorithm [20] was used to detect the lesions in MATLAB R2019a. Two individuals (without radiological expertise) were briefly trained by an expert to visualize the lesions in the affected lungs. The non-expert individuals provided the initial seed points in the respective foreground (all lesions) and background (rest of the image) and the lesion boundary was detected.
c. Refinement- The detected lesions were refined by morphological opening to remove the co-detected smaller/thinner bronchial structures and pulmonary vessels in the lung tissue.

## 2.4 Segmentation

a. A binary mask was created from the resultant detected lesions (expert and non-expert) and was used to segment the lesions from the 8-bit gray scale images.
b. The performance of the computer-aided segmentation was validated against the expert labelled lesion area using the Dice similarity coefficient (DSC) given by $DSC = \frac{2TP}{2TP+FP+FN}$, where TP= True Positive, FP=False Positive. and FN= False Negative
c. Each connected component was considered as a single lesion. Feature extraction was performed for all the segmented lesions.

## 2.5 Feature Extraction and Quantification

Since in CT, the bone has the highest CT value i.e. $\sim$ 400-1000 Hounsfield units (HU) and the air has the lowest i.e. -1000 HU, the vertebral disk (cancellous region) was taken as the maxima bone reference ($B$) and the air-region exterior to the chest in the same image was taken as the minima air reference ($A$) (Figure 2 colored box). The choice of the vertebral cancellous region was to reduce the CT-value variability within bones. The non-expert after segmenting the lesions gets automatically prompted to position a 30 pixel diameter circle each for the minima (air cavity) and maxima (vertebra) from the same CT image. The program then evaluates the $L_{norm}$. First, the mean gray intensities of $A$, $B$, $L$ were calculated from the region using $\{A, B, L\} = (1/n) \times (\sum_{i=1}^{n} x_i)$, where $x$ is the pixel intensity between 0 to 255. Once all the three values ($A$, $B$, $L$) are determined, the $L_{norm}$ was calculated by

$$L_{norm} = 100 \times \frac{(L - A)}{(B - A)} \quad | \quad 0 \leq L_{norm} \leq 100 \tag{1}$$

All analyses were performed in MATLAB R2019a. In case of multiple lesions, up to two lesions with the highest $L_{norm}$ were selected for severity assessment. Based on the values of $L_{norm}$, the lesion was categorized into three severity classes (S-1, S-2, S-3) using cut-off values evaluated from ROC analyses (discussed in subsequent section).

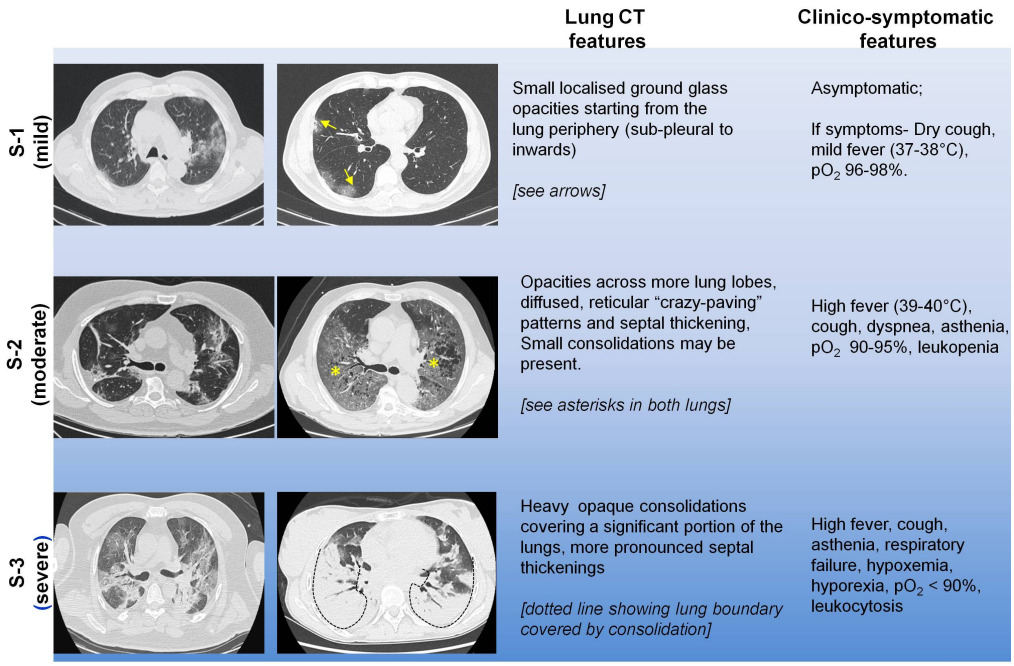


Figure 3: **Overview of clinical and chest CT features of pneumonia in COVID-19 patients** Columns 1 and 2 are representative expert assigned chest CT images for the three severity groups. C olumns 3 and 4 are the associated lung-CT features and clinico-symptomatic features. This is a guide to better understand the basis of expert classification (ground truth). *Note:* Features are an approximate guide based on previous literature [13, 15] and does not always correlate completely with severity.

## 2.6 Validation

### 2.6.1 Statistical Analysis

Pearson's correlation test was employed to evaluate the correlation between $L_{norm}$ and the mean gray scale intensity of the lesion (L). A total of 163 evaluations from both groups were considered, and the test was performed with two-tailed t-test and 95% confidence interval (CI).

One way ANOVA (with multiple comparisons) was performed to evaluate how well the feature $L_{norm}$ can delineate the three severity states and the p-value was estimated at 95% confidence interval. A two-tailed t-test was additionally performed to evaluate separation between group pairs.

To identify mild from non-mild cases and severe from non-severe cases, Receiver Operating Characteristic (ROC) curve analysis was performed at 95% confidence interval. The area under the curve as well as the optimum cut-off with the highest combination of sensitivity and specificity was determined. The radiological Lung-CT scores was the ground truth for allocating the individual groups to be delineated.

To evaluate the agreement between and within raters (1 radiology expert and 2 non-experts), $\kappa$-statistic was used (at 95% CI). All statistical analyses were performed in GraphPad Prism platform.

### 2.6.2 Numerical weighted accuracy

The evaluated results of quantified severity assessment using $L_{norm}$ by an expert and two non-experts was validated against the ground truth disease severity. To obtain the overall weighted percentage accuracy of agreement between the ground truth and $L_{norm}$ findings, we used:

$$Accuracy(weighted) = \left(1 - \frac{\sum_{i=1}^{n}|r-t|}{2 \times n}\right) \times 100\% \quad | \quad r, t = \{1, 2, 3\} \tag{2}$$

where $n$ is the number of cases, $r$= ground truth based severity, $t$=$L_{norm}$ based severity; 1, 2, 3 correspond to mild, moderate and severe classes.

### 2.6.3 Machine Learning for Estimating Severity

To evaluate the computational performance of $L_{norm}$ to achieve the three-class classification (mild, moderate, and severe), a number of machine learning based classifiers were employed e.g. Decision trees, Naïve Bayes, KNN, Ensemble classifiers etc. $L_{norm}$ values evaluated from multiple lesions by one expert and two non-experts which constituted a total of N=248 for classification. For classification, the sample size was partitioned randomly in a ratio of 60:40 (training:testing). A number of intensity (gray-scale intensity, lesion standard deviation) and Gray level co-occurence matrix (GLCM) texture features (angular moment, contrast, correlation, inverse difference, and entropy) were measured for the same set of lesions to compare their classification performance with $L_{norm}$. The classifiers were trained in MATLAB R2019a and were 10-fold cross validated in order to avoid over fitting by the classification models. Principle component analysis (PCA) was used for dimensional reduction in multivariate trained models with intensity and texture features. The testing accuracy of the best classification model was then determined.

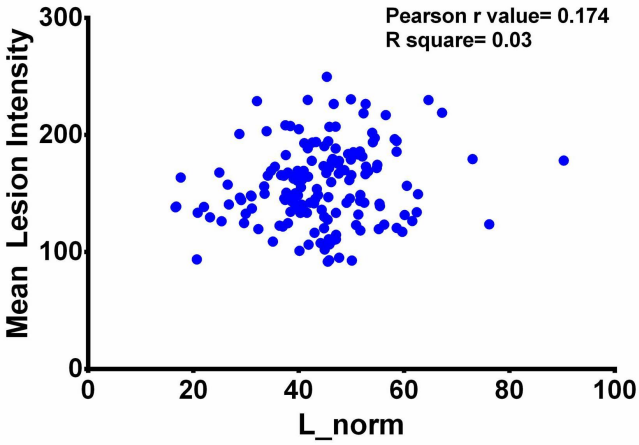


Figure 4: **Correlation between lesion gray scale intensity and $L_{norm}$.** Pearson's correlation test showing that the mean intensity of the lesions has almost no correlation with $L_{norm}$ (N=163).

## 3 Results

The CT image dataset presented key images, patient history, clinical, and radiological findings. After processing the entire dataset and filtering cases as per the inclusion/exclusion criteria, we determined that the images used in this study were procured from 27 different imaging centers spread across Italy (Supplementary Table S2). The severity index —S-1, S-2, and S-3 were assigned to the severity conditions—mild, moderate, and severe (Fig. 3, Supplementary Table S1). Since multi-center CT data was handle with a high variation between images we showed that the $L_{norm}$ has almost no correlation with the primary variable it is derived from i.e. the mean lesion intensity (Figure. 4).

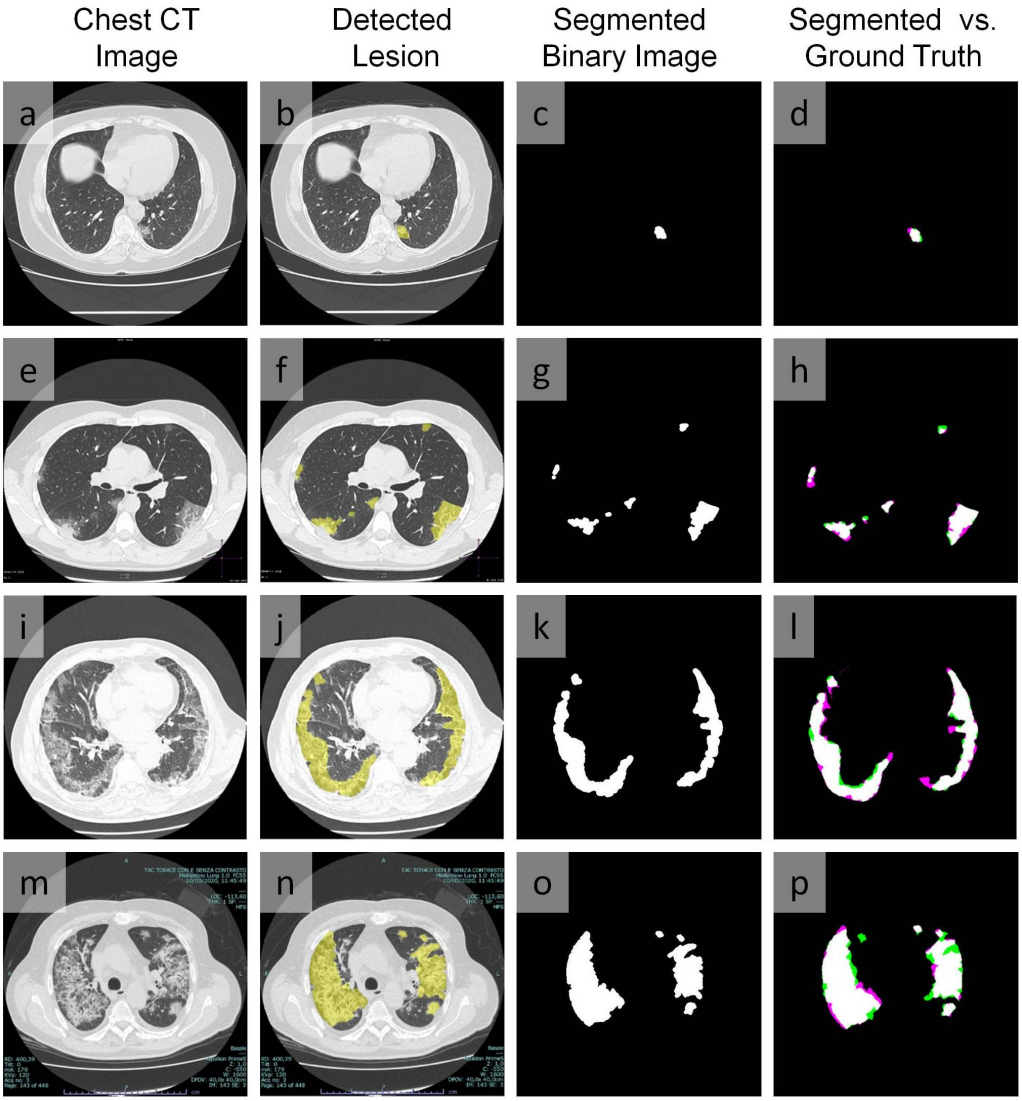


Figure 5: **Segmentation of lesions of pneumonia-affected lungs in COVID-19 patients.** (a,e,i,m) original chest CT images with lesions of different size, number, and localization, (b,f,j,n) lesions detected by the graph-cut based lesion detection, (c,g,k,o) binary images of the segmented lesions, (d,h,l,p) overlay of segmented result and expert-marked lesion area. The white area depict fully overlapped region, the orange and magenta areas depict exclusively segmented and expert-labeled regions respectively.

### 3.1 Lesion Detection and Feature Extraction

The lesions were detected using graph-cut region growing initiated by user-defined seed points for foreground and background followed by morphological opening (Fig. 5). We observed that the segmentation method detected lesions that are —small and faint (Fig. 5a-c), multiple with size variations

(Fig. 5e-g), periphery localized (Fig. 5i-k) as well as pan-lung (Fig. 5m-o) with significant overlap with expert-determined lesion area (Fig. 5d,h,l,p). A few more results to illustrate the segmentation performance of CT images from multiple centers are shown individually for mild, moderate and severe cases in Supplementary Fig. S1, S2, S3. A Dice similarity coefficient (DSC) of $0.887 \pm 0.05$ $(N = 30)$ was found when the segmentation results were compared against expert-marked lesion area.

### 3.2 Determination of Thresholds to Delineate the Severity Conditions

$L_{norm}$ feature values to identify the severity was determined using Equation 1 from segmented lesions shown in Fig. 2. The ground truth score determined by the in-house experts after going through the database reports (see general guidelines in Fig. 3) along with the evaluated $L_{norm}$ values are given in Supplementary Table S1. The distribution of $L_{norm}$ values in each of the ground truth categorized severity groups from the heuristic inputs of both experts (Fig. 6a) and non-experts (Fig. 6d) shows the features are statistically distinct for the three severity stages (p-value $< 0.0001$, 95% CI). Further, the optimum cut-off values of $L_{norm}$ to identify the three stages were determined by the ROC analysis (Fig. 6b,c,e,f). Based on the best sensitivity and specificity combinations from the ROC analysis (from both expert and non-expert data) to delineate severe from non-severe (sensitivity/specificity - 97.7%/97.49%) and mild from non-mild cases (sensitivity/specificity - 70%/87.88%) we determined the cut-off for each of the three categories (Table. 1). Both expert and non-expert assigned demarcations had a high accuracy for delineating severe from non-severe cases with ROC area under the curve (AUC) of 0.96 and 0.99 (Fig. 6 b,e). However, expert annotations were superior (AUC=0.96) for delineating mild from non-mild cases compared to non-expert (AUC=0.86) (Fig. 6 c,f). Thus, in conclusion, the feature $L_{norm}$ is found to be well-differentiated for the three severity conditions in a multi-center data across experts and non-experts.

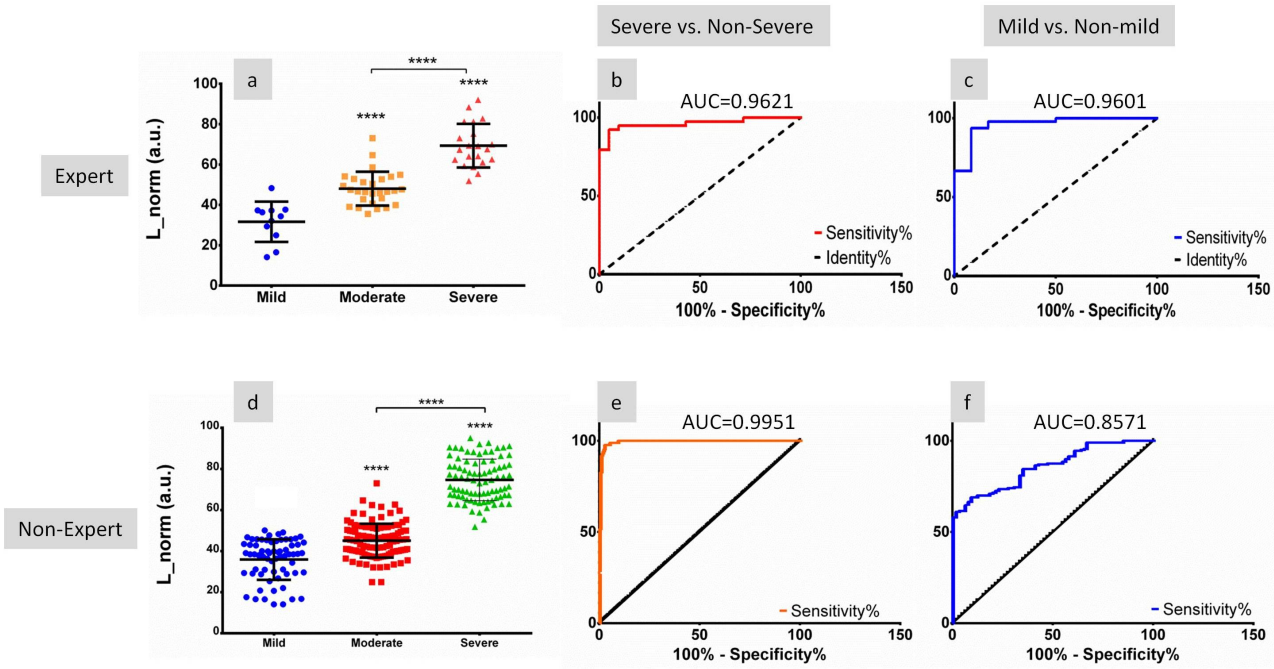


Figure 6: **Statistical analyses for determining $L_{norm}$'s potential in severity assessment.** (a-c) are statistical findings from one expert annotations (N=60) while (d-f) are from two non-expert annotations (N=180, up to 2 lesions per image). (a,d) shows $L_{norm}$ is well separated for the three severity cases,between all groups the p-value was <0.0001, (b,e) ROC analyses for classifying severe from non-severe class, (c,f) ROC analyses for classifying mild from non-mild class.

Table 1: Normalized lesion intensity $L_{norm}$ for each severity class

| Severity Score | $L_{norm}$ |
|---|---|
| S-1 (mild) | $< 45.80$ |
| S-2 (moderate) | 45.80-58.66 |
| S-3 (severe) | $>58.66$ |

Table 2: $\kappa$-statistic to evaluate the agreement of expert and non-expert raters to assign one of the 3 severity classes (mild, moderate, and severe) based on ROC-derived optimum $L_{norm}$ threshold.

| **Agreement Type** | **N** | **%agreement** | **$\kappa$-score** | **Remark [21]** |
|---|---|---|---|---|
| Intra-rater (non-exp1) | 30 | 96.67 | **0.965** | **almost perfect** |
| Intra-rater (non-exp2) | 30 | 96.67 | **0.965** | **almost perfect** |
| Inter-rater (non-exp1vs.2) | 30 | 83.33 | **0.825** | **almost perfect** |
| Inter-rater (exp vs. non-exp1) | 30 | 80 | **0.791** | **substantial** |
| Inter-rater (exp vs. non-exp2 | 30 | 86.67 | **0.859** | **almost perfect** |

### 3.3 Agreement Among Human Raters for Severity Assessment

To determine the accuracy of chest CT feature $L_{norm}$ based severity to match the annotated ground truth we evaluated the weighted accuracy as per equation 2. The accuracy was measured using the expert derived $L_{norm}$ to assess the closeness of the radiological (expert) measure to a multi-factor ground truth (symptoms, clinical findings, radiological findings etc.). The severity classes were assigned based on the thresholds determined in the previous section (Table 1). The weighted accuracy for the 3-class severity classification was determined to be 86.88%.

To further determine the agreement between raters and dependency of expertise levels and segmentation method, the statistical $\kappa$-scoring was done involving two non-experts and one expert (Table 2). Both the non-expert intra-rater agreements were the highest ($\kappa$ score 0.965) showing the fidelity of the segmentation method used. The agreement between expert and non-expert ranged from substantial to almost perfect [21]. In conclusion, the high agreement between and within raters demonstrates the reliability and reproducibility of the feature $L_{norm}$.

### 3.4 Machine Learning for evaluating classification accuracy

To establish that $L_{norm}$ can effectively classify severity in three classes without exclusively assigning optimized cut-offs (like from ROC analysis), we employed machine learning classifiers. A number of classification models were used for training with $L_{norm}$ values (Table 3). Additionally, we used different standard features of image gray-level intensity and texture associated with the chest CT lesion images to compare their classification performance with $L_{norm}$. Severity classes (S-1, S-2, and S-3) were assigned as per the annotated ground truth associated with the individual cases for training the models. It was found that the feature $L_{norm}$ had the highest classification accuracy among all the other individual as well as compound (multi-variate) features for all the classification models (88.2%). Among the classification models, the Decision Tree was determined to have the highest classification accuracy, followed by KNN and ensemble-learning models (Boosted and bagged trees). In conclusion, $L_{norm}$ alone has a superior 3-class classification performance as compared to the standard intensity and texture features.

Table 3: Comparison of classification accuracy based on various popular classification models to achieve the 3-class classification using different chest CT image features vs. $L_{norm}$

| | | **Model accuracy (%) of 3-class classification from different chest CT lesion features** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Classifier Type** | MI | MI+SD | AM | GLCM Contrast | GLCM Correlation | ID | GLCM Entropy | GLCM (All) | MI+SD+ GLCM | $L_{norm}$ |
| 1 | Decisison Tree | 66.3 | 67.5 | 44.1 | 51.6 | 48.9 | 48.4 | 44.1 | 52.2 | 50.5 | **88.2** |
| 2 | L-DA | 65.0 | 65 | 38.2 | 41.9 | 40.3 | 40.9 | 38.2 | 41.9 | 40.9 | **76.3** |
| 3 | Q-DA | 65.6 | 63.8 | 33.3 | 40.9 | 44.1 | 30.1 | 38.2 | 40.9 | 41.9 | **77.4** |
| 4 | Naive Bayes | 65.6 | 62.6 | 41.4 | 40.9 | 44.2 | 47.3 | 44.1 | 42.5 | 41.9 | **80.1** |
| 5 | Linear SVM | 63.2 | 66.3 | 35.5 | 39.2 | 37.1 | 35.5 | 34.4 | 39.2 | 39.2 | **78.5** |
| 6 | Quadratic SVM | 57.7 | 73.0 | 31.7 | 36.6 | 39.2 | 34.9 | 35.5 | 34.9 | 37.6 | **74.7** |
| 7 | Gaussian SVM | 66.9 | 73 | 40.3 | 38.7 | 40.3 | 45.2 | 45.7 | 38.2 | 38.1 | **83.3** |
| 8 | KNN | 63.8 | 73 | 46.2 | 50.5 | 48.4 | 50.5 | 49.7 | 49.5 | 48.9 | **86** |
| 9 | Cubic KNN | 63.8 | 66.3 | 41.9 | 47.8 | 41.9 | 43.5 | 43 | 47.3 | 46.8 | **86** |
| 10 | Weighted KNN | 62 | 71.8 | 45.2 | 51.1 | 48.4 | 50 | 48.9 | 50.5 | 48.4 | **81.7** |
| 11 | Boosted Trees | 61.3 | 70.6 | 43 | 51.6 | 50.5 | 53.2 | 45.7 | 51.1 | 48.9 | **82.8** |
| 12 | Bagged Trees | 62 | 71.2 | 43.5 | 51.1 | 49.5 | 52.2 | 47.8 | 49.5 | 48.4 | **83.3** |
| 13 | RUS Boosted Trees | 60.1 | 68.7 | 41.9 | 41.9 | 46.8 | 51.1 | 45.2 | 45.7 | 44.1 | **83.9** |

*Abbreviations: L-DA: Linear Discriminant Analysis, Q-DA: Quadratic Discriminant Analysis, SVM: Support Vector Machine, KNN: K-Nearest Neighbor, RUS: Random Undersampling, MI: Mean intensity, SD: Standard deviation, GLCM: gray level co-occurrence matrix, AM: GLCM angular moment; ID: GLCM Inverse Difference*

### 3.5 3-class vs. 2-class classification

In many scenarios delineation of severe to critical cases are more relevant and therefore we additionally showed the performance of $L_{norm}$ for classifying severe from the non-severe class. Upon investigating the results of 3-class classification models (Fig.7 a-c) it was seen that most of the errors in the model were concentrated in the mild and moderate classification (Fig.7 a,c). The confusion matrix of the trained Decision Tree model (Fig.7 c) revealed that none of the severe cases were mis-classified as mild or moderate. Thus we re-trained a decision tree model with only two classes i.e. severe and non-severe(Fig.7 d-f). Both mild and moderate cases (S-1 and S-2) were considered as non-severe here. This two-class classification showed a much higher AUC (0.99) and accuracy as compared to the 3-class classification (training (N=148) 98.9%, testing (N=100) 100%) . Additionally, the summary of the model accuracy for the 2-class classification for different trained models is summarized in Table 4 highlighting the enhanced performance compared to the 3-class classification.

Table 4: Accuracy of classification models to achieve the 2-class classification (severe and non-severe) using $L_{norm}$.

| | Decision Tree | L-DA | Q-DA | Naive Bayes | SVM | KNN | Bagged Trees |
|---|---|---|---|---|---|---|---|
| % accuracy | **98.9** | **98.4** | **98.4** | **98.9** | **98.4** | **98.9** | **98.9** |

## 4 Discussions

The emergence of CT imaging of lungs as an important tool for identifying severity, motivated the quest of a single feature that can be used across multiple imaging centers to perform severity classifications. The images in the dataset were captured from 27 different radiology units and expressed high variations in image quality and contrast (Supplementary Fig. S1,S2,S3). A number of imaging parameters (discussed previously) determine the intensity of the visualized lesion. Further, contrast adjustment during imaging is a common practice among radiologists to make distinct identification of affected lung regions, and in many practical situations, raw images may not be available or retrievable. To this aim, the feature $L_{norm}$ does not require raw or pre-processing of the enhanced images. This is essentially due to the inherent anatomical normalization with normalizing elements depending only

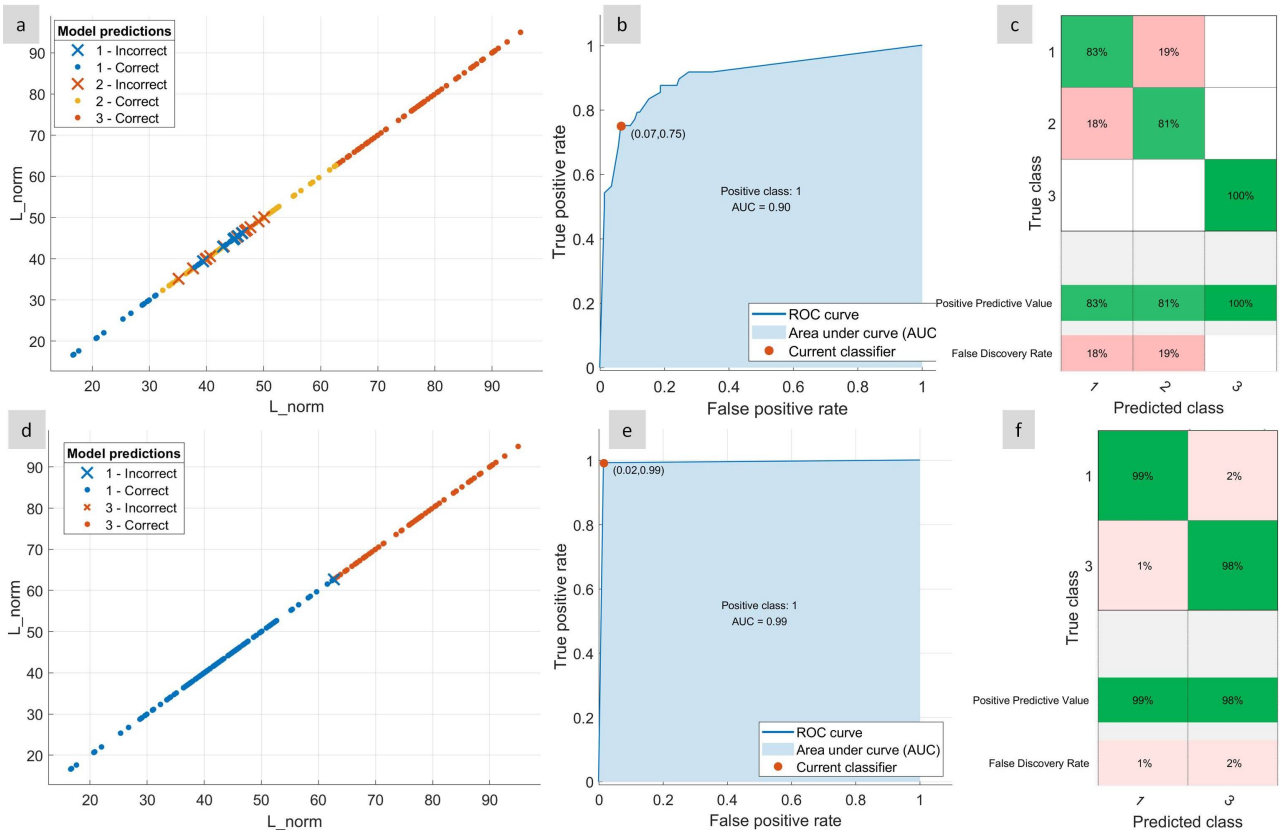


Figure 7: **Evaluation of $L_{norm}$ for 3-class and a 2-class (severe and non-severe) classification using a Decision Tree model.** (a) Classification performance for a 3-class 1. mild, 2. moderate, and 3. severe classification (training N=150), (b) corresponding ROC curve and (c) confusion matrix showing no mis-classification for severe cases, (d) performance of 2 -class classification 1. non-severe and 2. severe, (e) corresponding ROC curve and (f) confusion matrix showing the positive predictive value and false discovery rate.

on the same image-slice as the lesion. This ensures the reproducibility of the method across a range of CT units.

In the recent publications it was shown that mean lesion intensity does not have a very high correlation with the severity features of the lungs [13]. However, although $L_{norm}$ is primarily derived from the same lesion intensity it has a very higher performance in classifying severity. This is because $L_{norm}$ normalizes large variations among images and is independent of the post-processing and imaging parameters. The amount of this variation is clear with almost negligible correlation between mean lesion intensity and $L_{norm}$ (Figure 4).

The increased clinical burden of COVID-19 has expedited the necessity of assessing severity of individuals affected with the virus and allocation of radiological expert can be challenging. With a high agreement in severity assessment between the experts and non-experts, the method can be implemented by non-expert staffs with little training for routine evaluation of severity. The feature $L_{norm}$ shows almost similar performance in delineating severe cases from non-severe ones by both expert and non-expert annotations(Fig. 6). However, in dealing with delineating mild from non-mild cases expert based measurements provided a slightly higher accuracy over in non-expert based results (Fig. 6). Therefore, the expert dependency need to be modulated on requirement basis.

The chest CT gray scale intensity clinically translates to the increase in pathological deposition of exudates along with tissue involvement [16]. As $L_{norm}$ captures the chest CT intensity, its increase linearly correlates with the increase in disease severity. Besides, the linearity of $L_{norm}$ reduces the co-dependence of other image features and multi-variate classifier training in order to achieve better

classification results. Although specific cut-offs have been provided for classification, it needs to be noted that severity is more continuous than a discrete class bounded by a cut-off margin. Thus, the linear dependence of the $L_{norm}$ values with the disease severity makes interpretation of the disease condition easier and non-discrete. This is also the reason why even for weak classification learners like Decision Tree, the performance of the $L_{norm}$ is the highest. This reduces the computational complexity as well.

The different classification models that were employed to classify severity showed that the severity mis-classifications are rarely seen when identifying severe cases in the 3 class-classification. Furthermore, no cases were mis-classified more than one level of severity i.e. mild cases can be mis-classified as moderate but never as severe. This not only shows the fidelity of the feature but ensures almost no severe cases to be mis-classified. This became apparent when the classification models were trained for a 2-class severe vs non-severe classification (Fig. 7). The 2-class classification not only showed an extremely high AUC (0.99) but a very high classification accuracy across all the classifiers (Table 4).

Since the method is at this point semi-automated, it will be more desirable to achieve full automation by incorporating methods such as Deep Learning trained across data from a multitude of imaging centers with a wide range of imaging parameters.

The demonstrated quantitative severity assessment from chest CT in COVID-19 positive individuals, if implemented properly can help in managing the patients and provide the necessary treatment to reduce mortality and side-effects. Here we have outlined a scheme of general clinical work flow to demonstrate how the patient management can be done to incorporate the method (Figure 8).

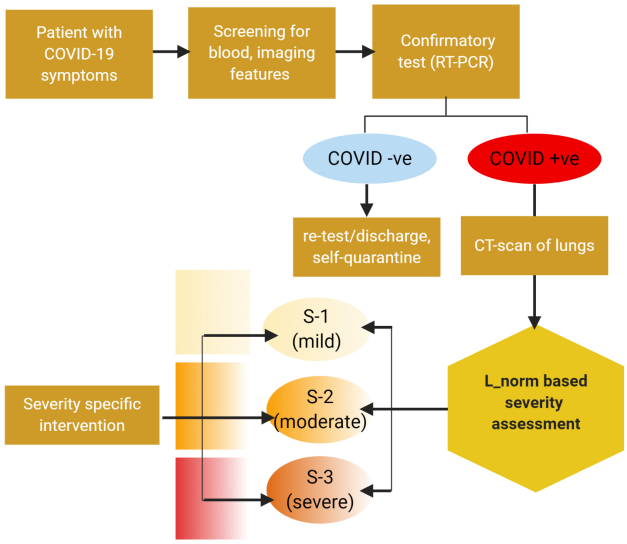


Figure 8: A proposed workflow of resource utilization and incorporation of severity assessment in diagnosis to therapy pipeline of COVID-19 disease in clinical settings.

## 5 Conclusion

To summarize, the article illustrates the lung-CT feature of COVID-19 patients to evaluate their severity quantitatively and group them in three classes of severity —S-1(mild), S-2(moderate), and S-3(severe) using the feature $L_{norm}$. This feature helps identification of severity groups which can help in therapeutic decision making for reducing risks and mortality in such a wide pandemic. The simplicity of the method along with high agreement score makes it a potential tool to be incorporated in the clinical diagnosis-therapy pipeline for management of COVID-19.

## References


[1] P. Kakodkar, N. Kaka, and M. Baig, "A comprehensive literature review on the clinical presentation, and management of the pandemic coronavirus disease 2019 (covid-19)," *Cureus*, vol. 12, no. 4, 2020.

[2] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.

[3] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, and W. Ji, "Sensitivity of chest ct for covid-19: comparison to rt-pcr," *Radiology*, p. 200432, 2020.

[4] Q.-Y. Peng, X.-T. Wang, L.-N. Zhang, C. C. C. U. S. Group, *et al.*, "Findings of lung ultrasonography of novel corona virus pneumonia during the 2019–2020 epidemic," *Intensive Care Medicine*, p. 1, 2020.

[5] E. Poggiali, A. Dacrema, D. Bastoni, V. Tinelli, E. Demichele, P. Mateo Ramos, T. Marcianò, M. Silva, A. Vercelli, and A. Magnacavallo, "Can lung us help critical care clinicians in the early diagnosis of novel coronavirus (covid-19) pneumonia?," *Radiology*, p. 200847, 2020.

[6] J. P. Kanne, B. P. Little, J. H. Chung, B. M. Elicker, and L. H. Ketai, "Essentials for radiologists on covid-19: an update—radiology scientific expert panel," 2020.

[7] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, "Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: a descriptive study," *The Lancet Infectious Diseases*, 2020.

[8] Y.-Y. Zheng, Y.-T. Ma, J.-Y. Zhang, and X. Xie, "Covid-19 and the cardiovascular system," *Nature Reviews Cardiology*, pp. 1–2, 2020.

[9] C. Zhang, L. Shi, and F.-S. Wang, "Liver injury in covid-19: management and challenges," *The Lancet Gastroenterology & Hepatology*, 2020.

[10] Y. Cheng, R. Luo, K. Wang, M. Zhang, Z. Wang, L. Dong, J. Li, Y. Yao, S. Ge, and G. Xu, "Kidney disease is associated with in-hospital death of patients with covid-19," *Kidney international*, 2020.

[11] A. Zumla, J. F. Chan, E. I. Azhar, D. S. Hui, and K.-Y. Yuen, "Coronaviruses—drug discovery and therapeutic options," *Nature reviews Drug discovery*, vol. 15, no. 5, p. 327, 2016.

[12] R. Yang, X. Li, H. Liu, Y. Zhen, X. Zhang, Q. Xiong, Y. Luo, C. Gao, and W. Zeng, "Chest ct severity score: An imaging tool for assessing severe covid-19," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, p. e200047, 2020.

[13] C. Shen, N. Yu, S. Cai, J. Zhou, J. Sheng, K. Liu, H. Zhou, Y. Guo, and G. Niu, "Quantitative computed tomography analysis for stratifying the severity of coronavirus disease 2019," *Journal of Pharmaceutical Analysis*, 2020.

[14] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, and L. Xia, "Serial quantitative chest ct assessment of covid-19: Deep-learning approach," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, p. e200075, 2020.

[15] K. Li, J. Wu, F. Wu, D. Guo, L. Chen, Z. Fang, and C. Li, "The clinical and chest ct features associated with severe and critical covid-19 pneumonia.," *Investigative radiology*, 2020.

[16] J. Wu, X. Wu, W. Zeng, D. Guo, Z. Fang, L. Chen, H. Huang, and C. Li, "Chest ct findings in patients with coronavirus disease 2019 and its relationship with clinical features," *Investigative radiology*, vol. 55, no. 5, pp. 257–261, 2020.

[17] N. Çullu, S. Kalemci, Ö. Karakaş, İ. Eser, F. Yalçın, F. N. Boyacı, and E. Karakaş, "Efficacy of ct in diagnosis of transudates and exudates in patients with pleural effusion," *Diagnostic and Interventional Radiology*, vol. 20, no. 2, p. 116, 2014.

[18] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.

[19] F. Pan, T. Ye, P. Sun, S. Gui, B. Liang, L. Li, D. Zheng, J. Wang, R. L. Hesketh, L. Yang, *et al.*, "Time course of lung changes on chest ct during recovery from 2019 novel coronavirus (covid-19) pneumonia," *Radiology*, p. 200370, 2020.

[20] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *International journal of computer vision*, vol. 70, no. 2, pp. 109–131, 2006.

[21] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.