OPEN

# A Multiclass Radiomics Method–Based WHO Severity Scale for Improving COVID-19 Patient Assessment and Disease Characterization From CT Scans

*John Anderson Garcia Henao, PhD, Arno Depotter, MD, Danielle V. Bower, MD, PhD, Herkus Bajercius, MD, Plamena Teodosieva Todorova, MD, Hugo Saint-James, MD, Aurélie Pahud de Mortanges, MD, Maria Cecilia Barroso, MD, Jianchun He, PhD, Junlin Yang, MSc, Chenyu You, MSc, Lawrence H. Staib, PhD, Christopher Gange, MD, Roberta Eufrasia Ledda, MD, Caterina Caminiti, MD, Mario Silva, MD, PhD, Isabel Oliva Cortopassi, MD, Charles S. Dela Cruz, MD, PhD, Wolf Hautz, MD, Harald M. Bonel, MD, Nicola Sverzellati, MD, PhD, James S. Duncan, MD, Mauricio Reyes, PhD, and Alexander Poellinger, MD*



**Objectives:** The aim of this study was to evaluate the severity of COVID-19 patients' disease by comparing a multiclass lung lesion model to a single-class lung lesion model and radiologists' assessments in chest computed tomography scans.

**Materials and Methods:** The proposed method, AssessNet-19, was developed in 2 stages in this retrospective study. Four COVID-19–induced tissue lesions were manually segmented to train a 2D-U-Net network for a multiclass segmentation task followed by extensive extraction of radiomic features from the lung lesions. LASSO regression was used to reduce the feature set, and the XGBoost algorithm was trained to classify disease severity based on the World Health Organization Clinical Progression Scale. The model was evaluated using 2 multicenter cohorts: a development cohort of 145 COVID-19–positive patients from 3 centers to train and test the severity prediction model using manually segmented lung lesions. In addition, an evaluation set of 90 COVID-19–positive patients was collected from 2 centers to evaluate AssessNet-19 in a fully automated fashion.

**Results:** AssessNet-19 achieved an F1-score of 0.76 ± 0.02 for severity classification in the evaluation set, which was superior to the 3 expert thoracic radiologists (F1 = 0.63 ± 0.02) and the single-class lesion segmentation model (F1 = 0.64 ± 0.02). In addition, AssessNet-19 automated multiclass lesion segmentation obtained a mean Dice score of 0.70 for ground-glass opacity, 0.68 for consolidation, 0.65 for pleural effusion, and 0.30 for band-like structures compared with ground truth. Moreover, it achieved a high agreement with radiologists for quantifying disease extent with Cohen κ of 0.94, 0.92, and 0.95.

**Conclusions:** A novel artificial intelligence multiclass radiomics model including 4 lung lesions to assess disease severity based on the World Health Organization Clinical Progression Scale more accurately determines the severity of COVID-19 patients than a single-class model and radiologists' assessment.

**Key Words:** pulmonary disease, technology assessment, CT segmentation, radiomics modeling

(*Invest Radiol* 2023;00: 00–00)

Artificial intelligence (AI)–based lung image analysis models can optimize the identification of patients who need specialized care. Standardized intensive care unit admission criteria have been proven to safely reduce intensive care unit overload. However, state-of-the-art AI systems face challenges in standardizing COVID-19 severity states.[1–3]

A systematic review by Born et al[4] highlighted discrepancies between studies published by clinicians and AI communities on COVID-19 patient care. They found that most AI studies focused on diagnosis rather than tasks such as severity and prognosis assessment, which are more important in clinical practice. Also, it is pointed out that AI models have a low adoption rate in clinical settings due to the need for increased robustness and interpretability.[4] Deep learning approaches for automated COVID-19 diagnosis using medical images, or quantifying lung tissue involvement using computed tomography (CT) scans, have been proposed and have demonstrated potential.[5–11] However, these approaches currently face challenges in terms of standardization in patient condition characterization, making their implementation in health care systems difficult.[12]

On the other hand, AI models that assess COVID-19 patients' severity using medical imaging and clinical data must meet clinical requirements.[13–15] However, many existing approaches use a single-class lesion segmentation model, which only classifies voxels as "healthy lung" or "lesion," neglecting the various pathological patterns that occur during disease progression, reducing the models' accuracy in characterizing patient severity.

To overcome these issues, we propose AssessNet-19, an automated CT-based radiomics multiclass lung lesion segmentation model to assess disease severity based on a standardized World Health Organization

Received for publication April 7, 2023; and accepted for publication, after revision, May 26, 2023.

From the ARTORG Center for Biomedical Research, University of Bern, Bern, Switzerland (J.A.G.H., M.R.); Department of Diagnostic, Interventional, and Pediatric Radiology, Inselspital Bern, University of Bern, Bern, Switzerland (A.D., D.V.B., H.B., P.T.T., H.S.-J., M.C.B., H.M.B., A.P.); Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT (J.H., L.H.S., C.G., J.S.D.); Department of Biomedical Engineering, Yale University, New Haven, CT (J.H., J.Y., L.H.S., J.S.D.); Department of Electrical Engineering, Yale University, New Haven, CT (C.Y.); Section of "Scienze Radiologiche," Diagnostic Department, University Hospital of Parma, Parma, Italy (R.E.L., M.S., N.S.); Department of Medicine and Surgery, University of Parma, Italy (R.E.L., N.S.); Ricerca Clinica ed Epidemiologica, University Hospital of Parma, Parma, Italy (C.C.); Department of Radiology at Mayo Clinic College of Medicine and Science, Florida, Jacksonville, FL (I.O.C.); Section of Pulmonary, Critical Care, and Sleep Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT (C.S.D.C.); Department of Emergency Medicine, Inselspital University Hospital, University of Bern, Bern, Switzerland (W.H.); Campusradiologie, Department of Radiological Diagnostics, Lindenhofspital Bern, Bern, Switzerland (H.M.B.).; Campus Stiftung Lindenhof Bern, Bern, Switzerland (H.M.B.); and Department of Radiation Oncology, Inselspital, Bern University Hospital, Bern, Switzerland (M.R.).

Conflicts of interest and sources of funding: The study was funded by the Swiss National Science Foundation within the National Research Programme "COVID-19" (NRP 78) grant number 198388. The study was supported by Campus Stiftung Lindenhof Bern and the Swiss Institute for Translational and Entrepreneurial Medicine, Bern, Switzerland. The authors have no conflicts of interest to declare.

Correspondence to: John Anderson Garcia Henao, PhD, ARTORG Center for Biomedical Research, University of Bern, Murtenstrasse 50, CH-3008 Bern, Switzerland. E-mail: john.garciahenao@artorg.unibe.ch.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).



ISSN: 0020-9996/23/0000–0000
DOI: 10.1097/RLI.0000000000001005

Clinical Progression Scale (WHO-CPS) for COVID-19 patients.[16] We hypothesize that evaluating patient disease severity by considering various pathological lung imaging findings, such as ground-glass opacities (GGOs), consolidations (CONs), pleural effusion (PLE), and band-like structures (BANs), can improve accuracy and contribute to identifying radiological markers to characterize COVID-19 disease severity.



## MATERIALS AND METHODS

### Study Design

This study collected CT imaging exams and clinical data retrospectively from COVID-19 patients with acute lung disease from four medical centers: Inselspital Bern, University of Bern in Switzerland (IBE), Lindenhofspital Bern in Switzerland (SLB), University Hospital of Parma in Italy (UPA) and Yale University - New Haven Hospital in the USA (UYA). The data were collected between March 2020 and November 2021 from COVID-19 patients with acute lung disease. The clinical data were obtained during routine clinical workup and retrospectively collected and anonymized. The subjects included in the study had to have a positive COVID-19 PCR test and CT scan, with imaging and clinical data collected within 24 hours of each other for consistency.

Including these four sites was motivated by establishing a diverse dataset that would promote clinical consistency and mitigate potential biases associated with training on data from a single source. The selection of multiple sites resulted in a heterogeneous contribution, ensuring a broader representation of cases and enhancing the generalizability of the findings. The data set compiled for this study encompasses a comprehensive range of disease severities. It comprises CT scans obtained from 4 different manufacturers, using various reconstruction kernels, and includes scans conducted with and without intravenous contrast. It is important to note that the primary focus of this study was not to compare the performance of individual hospitals but rather to develop a robust model capable of running across different technical configurations in future applications.

### Data

The study was approved by the Ethics Commission of the Canton of Bern (ID: 2020-02614, ID: 2020-00954), the Ethics Committee at Yale University–New Haven Hospital (ID: 2000027839), and the Ethics Committee at the University Hospital of Parma (ID: 1398/2020/OSS/AOUPR). All patients in the study gave consent for their data to be used for research. We retrospectively collected patients' medical imaging and clinical data from which a subset of the available cases was selected using 3 criteria: patients had to have an acute COVID-19 infection, a CT scan taken within 15 days before and 60 days after a positive COVID-19 test, and clinical data available within ±12 hours of CT acquisition. This study confirmed the presence of the SARS-CoV-2 virus in all included patients by retrieving their positive PCR test results from the database at each hospital center. The PCR test procedures followed internal hospital protocols by established guidelines by health authorities, including the WHO and local health agencies, ensuring the reliability and accuracy of the results.

The data assembly, curation, and image ground truth labeling were completed in 3 steps, as shown in Figure S1 (Supplementary Material, http://links.lww.com/RLI/A833). Imaging characteristics used in the developing and evaluation cohorts are summarized in Table S1 (Supplementary Material, http://links.lww.com/RLI/A833). An initial U-Net (R231) model released by Hofmanninger et al[17] was used to automatically segment the left and right lungs from CT scans to create baseline lung segmentations. Finally, radiologists reviewed the automatically generated lung segmentations, making necessary corrections and manually segmenting each lung lesion according to the segmentation protocol.

### Disease Severity Labeling–Ground Truth

The WHO score was fully automated using the WHO scoring algorithm on raw clinical data from the IBE and UYA centers and manually obtaining data from the UPA center's medical records. All scores were based on the WHO-CPS.[16] The WHO score was calculated for subjects at centers IBE and UPA using clinical data within 24 hours centered around the CT examination. At UYA, clinical data were recorded daily and matched with the CT examination data from that same day. Manual review at centers IBE and UYA confirmed the automated scoring. Patients who died within 12 hours of CT were not included in the study.

The severity of the disease is evaluated based on 4 stages: ambulatory mild disease (symptomatic but not requiring hospitalization), hospitalized moderate disease (hospitalization with minimal treatment), hospitalized severe disease (hospitalization with noninvasive ventilation), and intubated critical disease (hospitalization requiring intubation, mechanical ventilation, and possibly organ failure). In addition, the AI model's evaluation involved categorizing the WHO scores into 3, 4, and 5 severity stages, which were grouped into 3 categories for assessment purposes (see Table 1).

In this study, we used hierarchical multilabel classification to group the WHO severity scores into coarser labels. This approach was adopted to address the larger label space inherent in individual WHO scores, in contrast to the relatively smaller label space associated with the multilabel severity group. Our evaluation focused on selecting the appropriate number and hierarchy of labels for grouping the severity scores, ensuring coherence in the classification process. For the 3-label hierarchy, we collapsed the WHO scale as follows: "mild ambulatory" (MA) encompassing scores 1 to 3, "hospitalized disease" (HD) covering scores 4 to 6, and "intubated critical disease" (IC) representing scores 7 to 9. In the 4-label hierarchy, we categorized patients into MA for scores 1 to 3, "hospitalized moderate disease" (HM) for scores 4 and 5, "hospitalized severe disease" (HS) for score 6, and IC for scores 7 to 9. Finally, the 5-label hierarchy involved the following groupings: MA for scores 1 to 3, HM for scores 4 and 5, HS for score 6, IC for scores 7 and 8, and "intubated critical disease plus organ failure" (IC+) for score 9. By examining the different hierarchical label configurations, we assessed the performance and coherence of the selected multilabel hierarchy in accurately representing the severity of COVID-19 patients based on the WHO scores.

### Medical Image Labeling–Ground Truth

Figure 1 illustrates 5 pathological CT findings used to train the multiclass lesion segmentation model. The data curation process included manual segmentation of 10 equidistant slices per subject, covering the lung from apex to the base, taking 2–6 hours, depending on the case complexity. The segmentation team consisted of 2 experienced radiologists (20 and 9 years of experience), 2 residents (2 years of experience), and 4 medical students trained by expert radiologists. At least 1 other team member reviewed segmentations to ensure quality. The lung and lesion segmentation followed the 2008 thoracic imaging definitions of the Fleischner Society.[18] The multiclass segmentation protocol ensured that each lung lesion segmentation remained within the boundaries of the lung segmentation. In addition, strict nonoverlapping criteria were enforced for different lesion classes. This was implemented due to the nature of the U-Net multiclass segmentation network, which was specifically designed to assign a single label to each voxel.

Labeling the 5 radiographic pathologies involved creating a coarse mask of contiguous lesions using a paint tool and then manually correcting and checking the segmentation using tools such as the erase tool and logical operator tool in 3D slicer. The segmentation of each pathological lung imaging finding was performed differently. For example, GGO, BAN, PLE, and TBR lesion segmentation labels were created in the lung window, whereas the soft tissue window (W: 350; L: 50) was used for CON segmentation. CON lesions were initially identified using the threshold tool to create a coarse mask for segmentation. Subsequently, manual correction of the CON label was performed in the lung window using the erase tool. If overlapping borders

**TABLE 1.** The WHO-CPS and 3 Groups to Assess the Disease Status of COVID-19 Patients

| | | Groups to Assess the Disease Status of Patients With COVID-19 | | |
|---|---|---|---|---|
| **WHO Clinical Progression Descriptor** | **WHO Score** | **3 States** | **4 States** | **5 States** |
| Uninfected; no viral RNA detected | 0 | — | — | — |
| Asymptomatic; viral RNA detected | 1 | Ambulatory mild disease | Ambulatory mild disease | Ambulatory mild disease |
| Symptomatic; independent | 2 | | | |
| Symptomatic; assistance needed | 3 | | | |
| Hospitalized; no oxygen therapy* | 4 | Hospitalized disease | Hospitalized moderate disease | Hospitalized moderate disease |
| Hospitalized; oxygen by mask or nasal prongs | 5 | | | |
| Hospitalized; oxygen by NIV or high flow | 6 | | Hospitalized severe disease | Hospitalized severe disease |
| Intubation and mechanical ventilation: $Po_2/Fio_2 \geq 150$ or $Spo_2/Fio_2 \geq 200$ | 7 | Intubated critical disease | Intubated critical disease | Intubated critical disease |
| Mechanical ventilation: $Po_2/Fio_2 < 150$ ($Spo_2/Fio_2 < 200$) or vasopressors | 8 | | | |
| Mechanical ventilation: $Po_2/Fio_2 < 150$ and vasopressors, dialysis, or ECMO | 9 | | | Intubated critical disease plus organ failure |
| Dead | 10 | — | — | — |

Note: This study did not evaluate symptoms in nonhospitalized patients, and therefore, we did not distinguish between scores 1, 2 and 3. The severity scoring was based according to the WHO working group guidelines,[16] using the parameters of viral detection, hospitalizations, use of low-flow oxygen (by nasal cannula), or high-flow nonintubated oxygenation (by high-flow nasal cannula or continuous or noninvasive positive airway pressure ventilation), intubation and mechanical ventilation, the oxygenation ratios based on $Spo_2/Fio_2$ or $Po_2/Fio_2$, the administration of vasopressors, requirement of dialysis or ECMO (extracorporeal membrane oxygenation), and death. To calculate the oxygenation ratios, $Fio_2$ is represented as a fraction (ie, 0.5 for 50% inhaled oxygen).

*If hospitalized for isolation only, record status as for ambulatory patient.

ECMO, extracorporeal membrane oxygenation; $Fio_2$, fraction of inspired oxygen; NIV, noninvasive ventilation; $Po_2$, partial pressure of oxygen; $Spo_2$, oxygen saturation.



with PLE were present, they were subtracted using the logical operator tool. All vessels within the CON area were included, whereas bronchi and BAN were excluded if not filled with fluid. After segmenting CON, the remaining opacified lung lesions were segmented as GGO. For GGO, the segmentation was carried out manually instead of using a threshold method. Large and intermediate vessels and visible bronchi were excluded. Band-like structures were defined as dense structures with a tubular-like shape, excluding pleura, and atelectasis. Three-dimensional visualization was used to identify BAN structures, as they tend to be intertwined with CON or GGO. The bronchial lumen was segmented for the TBR class, using 3D visualization to address motion artifacts and pseudobronchi. Please refer to the Supplementary Material, http://links.lww.com/RLI/A833, for visualizing the contouring process as per the segmentation protocol.

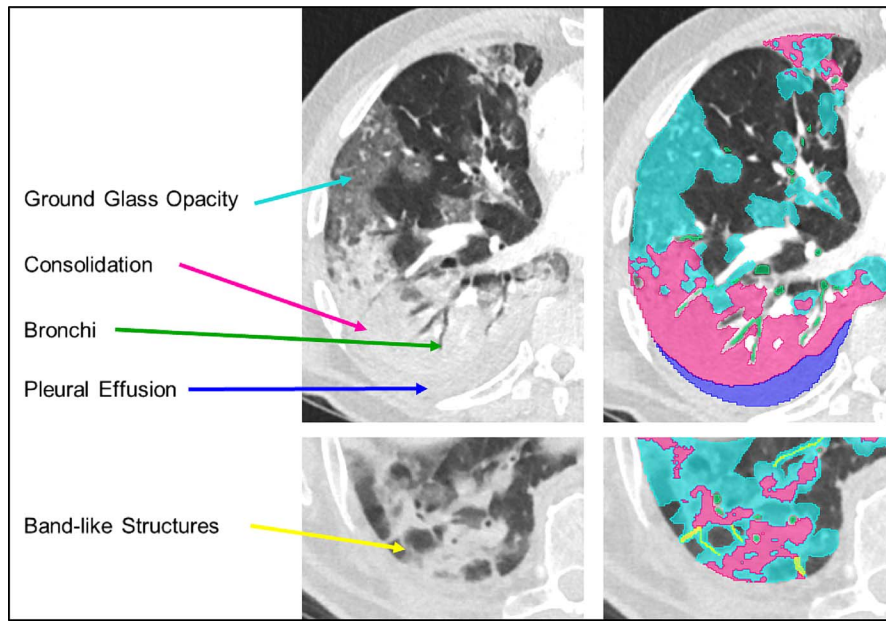


**FIGURE 1.** Manual segmentation of 5 different lesion classes such as ground-glass opacity (GGO), consolidation (CON), pleural effusion (PLE), band-like structure (BAN), and bronchi/traction bronchiectasis (TBR).

## Data-Centric AI Model to Automate the Multiclass Lesion Segmentation and Disease Severity Assessment

AssessNet-19 model is a data-centric AI model developed through incremental cycles, where subjects were selected from the needs of the previous model. Figure 2 shows the final design of the AssessNet-19 model.[19] The pipeline includes image preprocessing, lung segmentation, multiclass lesion segmentation, and radiomics feature extraction for severity assessment prediction.

The image preprocessing pipeline extracts axial slices and corresponding lesion segmentations from each CT scan and reshapes them to fit the 2D format required by the nnU-Net framework.[20] It also uses cropping, normalization, and resampling techniques, such as cropping to intensity values between 0.5th and 99.5th percentiles, normalizing using a $z$-score, and resampling to median voxel size using third-order spline interpolation for image data and nearest-neighbor interpolation for segmentation masks. The multiclass lung and lesion segmentation models use a 2D U-Net architecture,[21] and were implemented separately with the nnU-Net framework.[20] They were trained on 118 subjects using an NVIDIA-RTX-A6000, taking 16 hours for 1000 epochs per fold, averaging 57.90 ± 0.54 seconds per epoch for the lesion segmentation model, and 15.29 hours for 1000 epochs per fold for the lung segmentation model, averaging 55.06 ± 0.38 seconds per epoch. Section S1 in the Supplementary Material, http://links.lww.com/RLI/A833, provides the implementation details, and Figure S2 in the Supplementary Material, http://links.lww.com/RLI/A833, presents the learning curves for internal training and validation sets for each model. One hundred seven radiomics features were extracted from each axial slice per subject and each lesion class using the pyRadiomics library.[22] Essential features from each lesion were selected using LASSO.[23] Following the image biomarker standardization initiative, shape features were normalized based on lung segmentation to prevent bias due to lung anatomy.[24] Finally, the radiomics-based

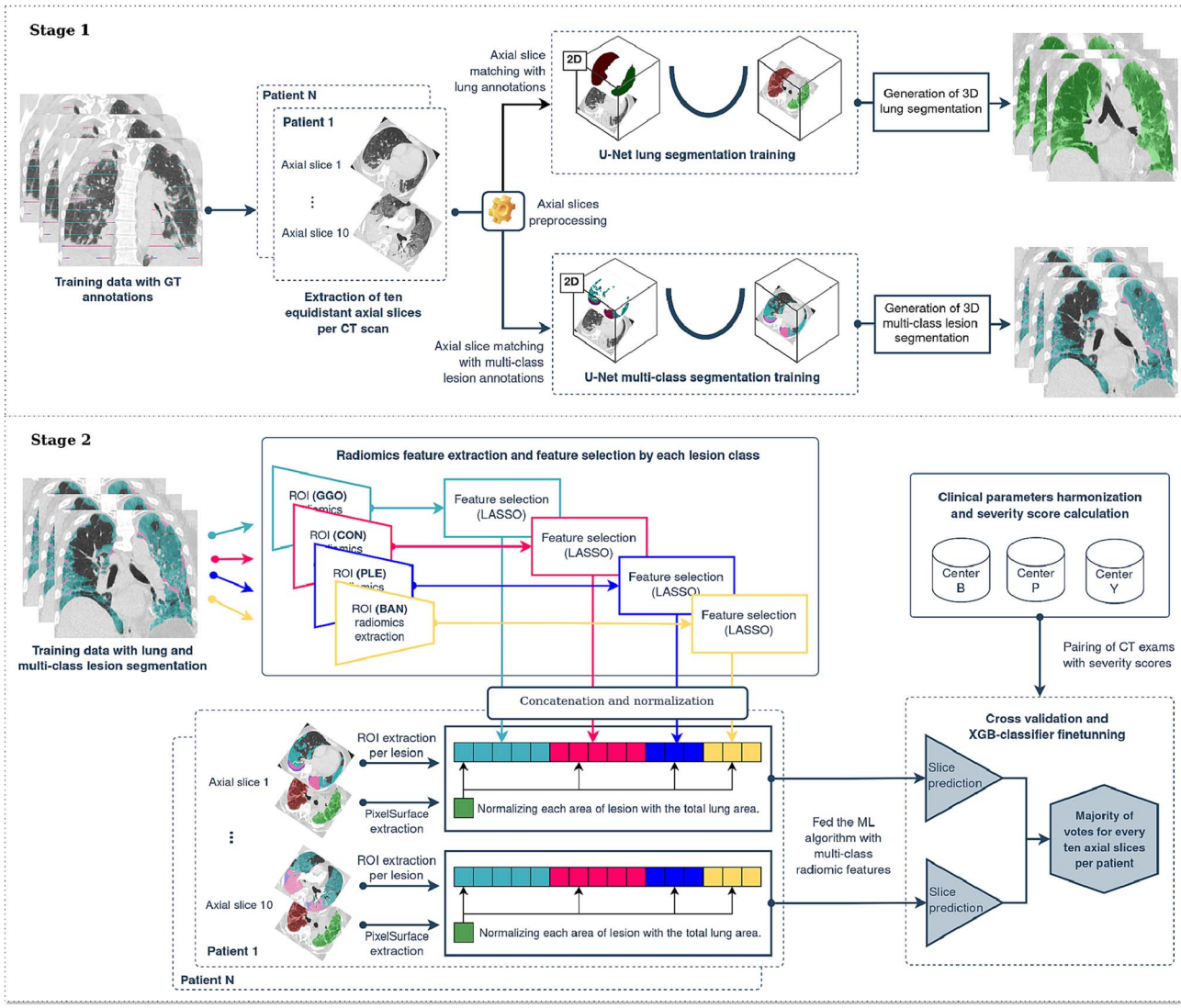


**FIGURE 2.** Overview of the AssessNet-19 model, a 2-stage pipeline for assessing COVID-19 patients' disease severity. First stage: Ten equidistant axial slices are extracted from each CT scan and paired with ground truth segmentation to train two 2D U-Net networks for lung and multiclass lesion segmentation. The 2D segmentation outputs are then used to construct the 3D volume of the lungs and multiclass lesions for quantification. Second stage: Radiomics feature extraction and selection process applied to each lesion class. Then, the elected features are concatenated, normalized, and inputted into the ML algorithm. Finally, the model was fine-tuned through cross-validation and XGB-classifier based on majority votes for every 10 axial slices per subject.

severity prediction model was trained using the extracted features, and various machine learning models were tested using F1-score as a metric on a 5-fold cross-validation procedure. The best-performing method was XGBoost,[25] which was chosen for evaluation in the test cohorts.

## Benchmarking

In the severity assessment by radiologists, 3 experienced lung radiologists with 20, 14, and 9 years of experience qualitatively and quantitatively assessed the disease severity using a 4-class severity scale and the disease extent for GGO and CON in the percentage of lung volume. The radiologists were blinded to all patient information, including the final severity score. Furthermore, we compared 3 ways of categorizing the disease severity as a 3-, 4-, or 5-label hierarchy to identify the suitable hierarchical multilabel classification task in terms of performance and coherence to represent the severity of disease states based on the -WHO-CPS.

## Statistical Analysis

The statistical analysis focused on evaluating the quality of automated lung and multiclass lesion segmentation using 2 metrics: Dice (Dice similarity coefficient) and Hausdorff distance. In addition, the performance of the WHO severity prediction model was assessed using multiple metrics, including confusion matrices for accuracy analysis, AUC-ROC (area under the receiver operating characteristic curve) for performance evaluation, and F1-score for a comprehensive assessment. These metrics and the corresponding confusion matrices provide insights into the model's effectiveness in accurately classifying disease severity categories.

# RESULTS

## Data Set Stratification and Patient Characteristics

The development and evaluation cohorts were compiled to ensure a balanced distribution of WHO scores. A stratified shuffle split approach was used to divide the development cohort into train and test sets, while preserving the same percentage for each WHO class as in the complete set. Figure 3 shows the distribution of WHO scores and disease severity labels in the training, testing, and second testing sets. First, a development cohort of 145 subjects: 70 from center IBE, 31 from center UPA, and 44 from center UYA. A total of 1450 axial slices were manually segmented. The subjects were then randomly divided into a set of 118 cases for training and a set of 27 cases for testing the AssessNet-19 model. The evaluation set comprised 90 subjects: 78 from IBE and 12 from UPA. This cohort was used to evaluate AssessNet-19 in a fully automated fashion. The study involved patients who were transferred from other hospitals in a more critical condition, which resulted in limited data availability regarding the timing of the first PCR test. The study included 58 CT scans conducted before a confirmative positive COVID-19 PCR test, representing a subset of 235 CT scans collected from 3 hospitals.

The first cohort was divided into training and testing sets using stratified sampling based on WHO score, deceased subjects, and CT kernels. Figure 3 shows the distribution of WHO scores for the development cohort and second test set. Table 2 summarizes the clinical characteristics recorded for each partition in the training and testing sets, including demographics, anthropometric variables, comorbidities, laboratory variables, and hospitalization characteristics obtained from medical records. This study used clinical data to calculate the WHO severity score, following the guidelines provided by Marshall et al.[16] To compute the WHO severity score, we derived the following clinical parameters from the available clinical variables: hospitalization status, mortality status, low $SpO_2$ levels, low $PO_2$ levels, vasopressor usage, intubation or tracheostomy procedure, high-flow oxygen therapy requirement, low-flow oxygen therapy requirement, dialysis requirement, and ECMO (extracorporeal membrane oxygenation) support.

## Multiclass Lesion Segmentation Performance

We evaluated the performance of AssessNet-19, our automated multiclass lung lesion segmentation model, using a test set of 27 subjects. AssessNet-19 was trained with 10 equidistant axial slices per CT scan, addressing the multiclass problem with standard nnU-Net hyperparameters. The learning curves are available in the Supplementary Material, http://links.lww.com/RLI/A833.

The evaluation demonstrated that AssessNet-19 consistently achieved accurate segmentations across all disease severities, aligning

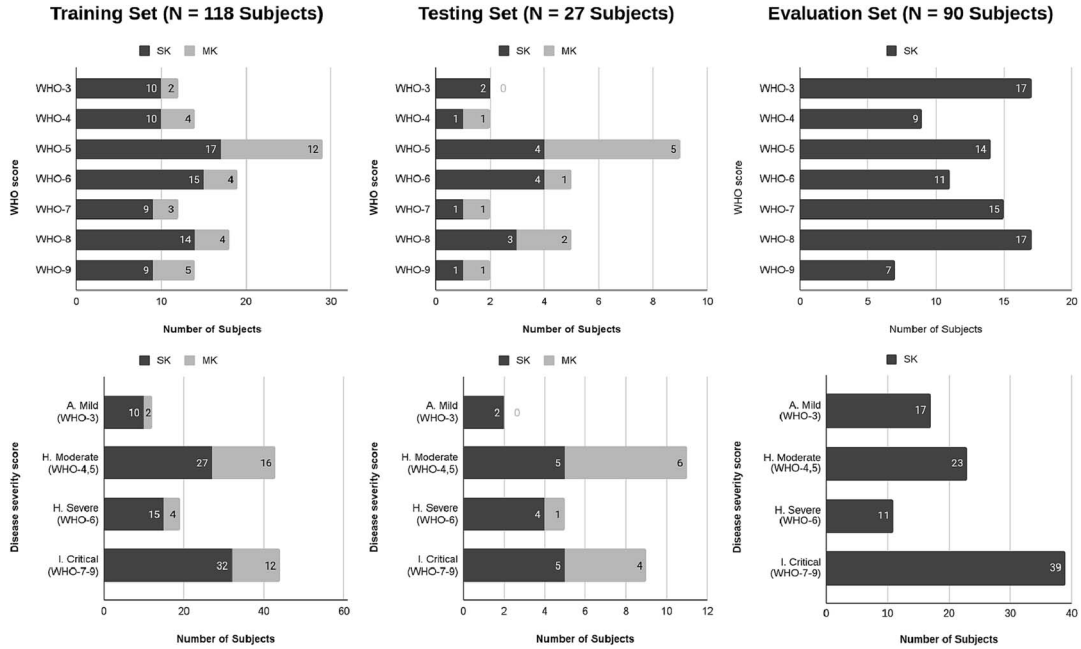


**FIGURE 3.** Distribution of WHO scores and disease severity labels among the training, testing, and evaluation sets. The abbreviations used in the figure are as follows: N represents the number of subjects, SK refers to soft-kernel, and MK represents medium-soft-kernel.

**TABLE 2.** Patient Characteristics Description Among the Training, Testing, and Second Testing Set

| Clinical Characteristics | Training Set n = 118 Distribution | IQR (25%, 75%) | Available | Testing Set n = 27 Distribution | IQR (25%, 75%) | Available | Second Testing Set n = 90 Distribution | IQR (25%, 75%) | Available |
|---|---|---|---|---|---|---|---|---|---|
| Demographic characteristics | | | | | | | | | |
| Age | 62.4 ± 14.3 | (54, 70) | 118 | 63.3 ± 14.78 | (54.5, 73.5) | 27 | 61.4 ± 11.97 | (54, 69) | 80 |
| Gender (female) | 41 (34.74%) | — | 118 | 9 (33.33%) | — | 27 | 24 (30.0%) | — | 80 |
| Gender (male) | 77 (65.25%) | — | 118 | 18 (66.66%) | — | 27 | 56 (70.0%) | — | 80 |
| Anthropometric characteristics | | | | | | | | | |
| Height, cm | 171.82 ± 9.47 | (−165, 178) | 72 | 167.04 ± 10.76 | (158.3, 172) | 15 | 171.11 ± 8.28 | (−165, 176) | 53 |
| Weight, kg | 85.69 ± 16.84 | (73.9, 95.9) | 79 | 86.30 ± 22.91 | (70.4, 99.6) | 17 | 83.84 ± 14.99 | (72.4, 92.2) | 52 |
| BMI | 29.29 ± 5.64 | (25.9, 31.9) | 73 | 31.09 ± 6.86 | (27.3, 36.3) | 16 | 29.04 ± 5.99 | (24.3, 34.4) | 47 |
| Comorbidities characteristics | | | | | | | | | |
| Asthma | 10 (8.69%) | — | 115 | 4 (14.81%) | — | 27 | 4 (5.0%) | — | 80 |
| Diabetes | 27 (24.32%) | — | 111 | 12 (48.0%) | — | 25 | 33 (41.25%) | — | 80 |
| COPD | 15 (13.51%) | — | 111 | 3 (12.%) | — | 25 | 15 (18.75%) | — | 80 |
| Lung fibrosis | 4 (3.47%) | — | 115 | 0 (0.0%) | — | 26 | 1 (1.25%) | — | 80 |
| Laboratory characteristics | | | | | | | | | |
| eGFR | 70.72 ± 26.79 | (47.0, 90.0) | 108 | 71.08 ± 29.48 | (46.7, 90.0) | 24 | 72.13 ± 23.90 | (58.5, 90.0) | 79 |
| WBCs | 9.04 ± 5.08 | (5.3, 11.0) | 84 | 9.93 ± 3.63 | (7.4, 13.9) | 17 | 11.10 ± 5.01 | (8.01, 23.2) | 45 |
| Lymphocytes | 1.13 ± 0.68 | (0.8, 1.3) | 64 | 1.03 ± 0.49 | (0.64, 1.31) | 16 | 1.06 ± 0.59 | (0.71, 1.43) | 16 |
| Neutrophils | 6.84 ± 4.28 | (3.9, 8.5) | 45 | 9.39 ± 3.76 | (6.55, 12.0) | 10 | 8.48 ± 6.46 | (3.3, 13.5) | 12 |
| CRP | 126.4 ± 100 | (36.9, 217) | 72 | 112.7 ± 113 | (38.8, 157) | 15 | 142.5 ± 107 | (62.5, 222) | 43 |

The clinical variables were obtained within ±12 hours of CT acquisition per subject.

Note: Continuous variables are represented as median, standard deviation, and interquartile range (IQR). Categorical variables are expressed as numbers and percentages of the available subjects.

BMI, body mass index; COPD, chronic obstructive pulmonary disease; eGFR, estimated glomerular filtration rate; WBCs, white blood cell count; CRP, C-reactive protein.



well with the ground truth. The model's performance varied across different lesion categories, with mean Dice similarity coefficients of 0.7 ± 0.27 for GGO, 0.68 ± 0.34 for CON, 0.65 ± 0.31 for PLE, and 0.30 ± 0.16 for BAN. In addition, AssessNet-19 exhibited improved consistency in segmenting shapes and sparse lesions, as indicated by the smaller Hausdorff distance. For a qualitative comparison, please refer to Figure 4, which illustrates segmentations produced by AssessNet-19 and the corresponding ground truth for selected cases involving COVID-19 patients with varying disease severities.

## Radiomics Signatures to Characterize the COVID-19 Disease

Radiomics signatures play a crucial role in characterizing COVID-19 disease, and our study used a comprehensive process to extract and analyze quantitative features from medical images. This process encompassed automated multiclass lesion segmentation to define regions of interest (ROIs), extraction of radiomic features related to shape, intensity, texture, and spatial relationships within the ROIs, and normalization of the radiomics features. The primary goal was to reduce the dimensionality of the feature space and identify the most relevant features for classification or prediction tasks.

In our study, we focused on identifying 4 key radiomics signatures that effectively characterize the severity of COVID-19 disease. These signatures provide valuable insights into the assessment task, specifically lung lesion segmentation and lesion extent quantification. Figure 5 visually presents the radiomics signatures for the 4 severity states: MA, HM, HS, and IC. Figure 5A showcases the radiomics signatures for the single-class model, whereas Figure 5B displays the signatures for the multiclass model. The spider charts in these figures illustrate the average values of the *z*-normalized radiomics features used in both models. Figures 5C and 5D demonstrate 3D reference lung and lesion segmentations, respectively, highlighting the segmentation outputs for each disease severity state from the single and multiclass models.

To provide further insights into the classification process, we present Table 4, which showcases the relationship between laboratory test results, CT scans of lung lesions, and radiomics signatures in classifying the 4 severity states of COVID-19. This table includes information on crucial laboratory tests (lymphocytes, neutrophils, white blood cell count [WBCsn], and estimated glomerular filtration rate [eGFR]) for each severity stage. The laboratory test results were obtained within ±12 hours of CT acquisition per subject. The table's CT lung lesion quantification section displays the extent of 4 types of lung lesions (GGO, CON, PLE, and BAN) for each severity stage, represented as percentages commonly used by radiologists in clinical practice. Finally, the radiomics disease signature section presents the features used in the multiclass model for each stage of COVID-19 severity. These features encompass lesion extension, intensity histograms, and various texture features such as the co-occurrence matrix, size zone matrix, neighboring tone difference matrix, dependence matrix, and run length matrix. This information provides additional insights into the patient's condition and aids in explaining the predictions made by the AssessNet-19 model.

Significant differences were identified ($P < 0.001$) between the 4 severity states for lymphocytes and eGFR. In contrast, little differences were observed ($P = 0.015$ and $P = 0.0012$) between the neutrophil disease states and WBCsn, respectively. In addition, significant differences were observed in CT of lung lesions quantification ($P < 0.001$) among the 4 COVID-19 severity states for CON, PLE, and BAN. In contrast, less pronounced differences were found ($P = 0.039$) between the 4 disease states for GGO. In some cases, patients in the severely hospitalized condition

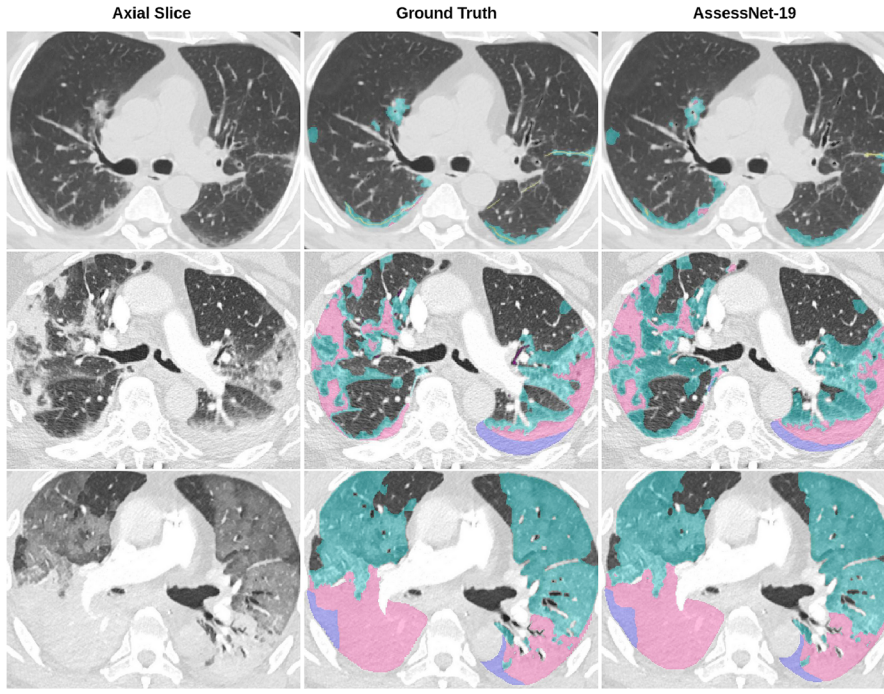


**FIGURE 4.** Qualitative results of AssessNet-19 for multiclass lesion segmentation in COVID-19 patients with varying disease severities, including ambulatory mild, hospitalized moderate, and critically intubated cases.

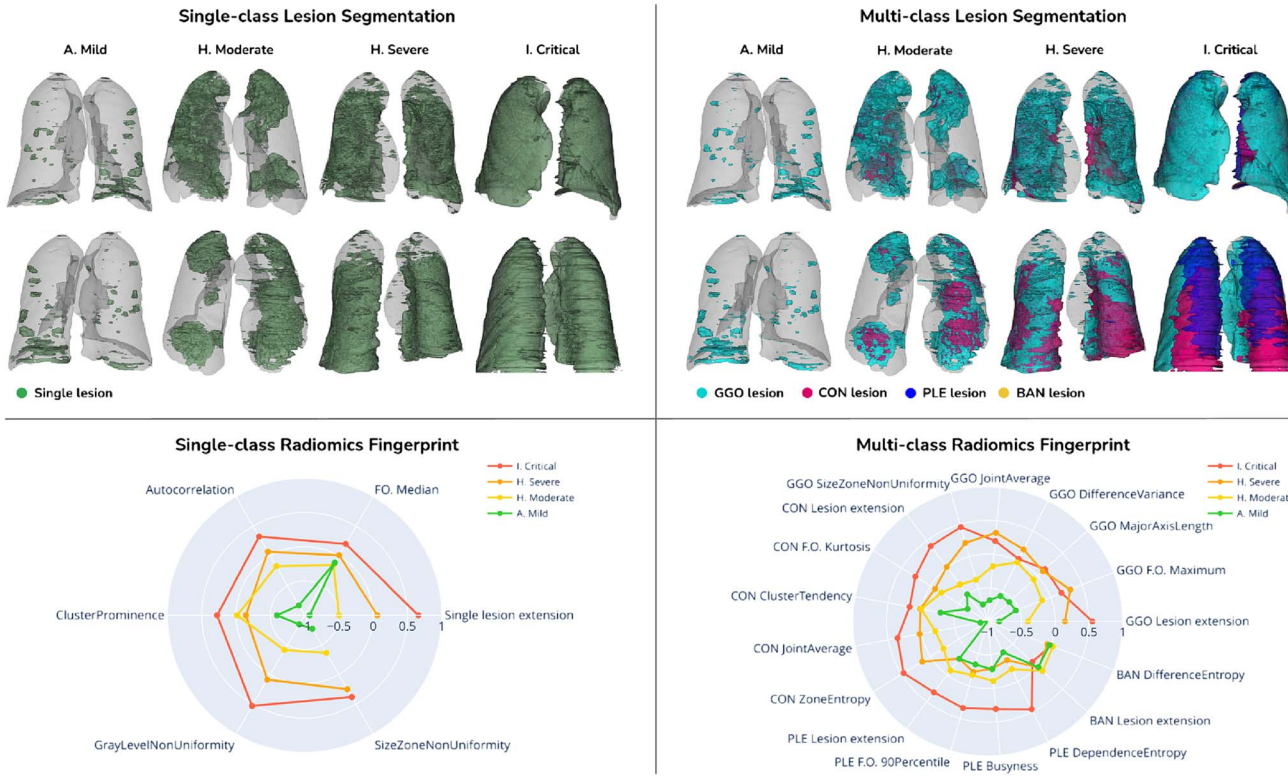


**FIGURE 5.** Radiomics signatures to characterize the COVID-19 disease. Radiomics signature of single-class and multiclass models using 4-disease state classification and a representative 3D lung and lesion segmentation for each disease state. The radiomics features were normalized and mainly composed of lesion extension, intensity histograms, and texture features such as the co-occurrence matrix, size zone matrix, neighboring tone difference matrix, dependence matrix, and run length matrix. Radiomics values fall within a range of –1 to 1.

**TABLE 3.** Quantitative Evaluation Among Single-Class Lesion Model, Multiclass Lesion Model, and Radiologists' Qualitative Score Assessment on the Development Cohort Using 27 Subjects Manually Segmented and the Second Evaluation Cohort Using 90 Subjects Fully Automated Segmented

| | Development Cohort–Ground Truth Test (n = 27) | | | Evaluation Test (n = 90) | | |
|---|---|---|---|---|---|---|
| | Single-Class Model | Multiclass Model | Radiologists' Quality Score | Single-Class Model | Multiclass Model | Radiologists' Quality Score |
| Classification Task | F1-Score | F1-Score | F1-Score | F1-Score | F1-Score | F1-Score |
| 3-WHO Classes | 0.71 ± 0.03 | **0.90 ± 0.03** | 0.63 ± 0.10 | 0.66 ± 0.01 | **0.79 ± 0.02** | 0.69 ± 0.03 |
| 4-WHO Classes | 0.52 ± 0.03 | **0.74 ± 0.02** | 0.45 ± 0.09 | 0.64 ± 0.02 | **0.76 ± 0.02** | 0.63 ± 0.02 |
| 5-WHO Classes | 0.39 ± 0.03 | **0.67 ± 0.03** | — | 0.51 ± 0.02 | **0.66 ± 0.01** | — |

Radiologists' Qualitative Score = the mean F1-score was determined by majority voting among 3 radiologist experts to assess the severity score qualitatively using only CT images.



**TABLE 4.** Radiomics Signatures to Characterize the COVID-19 Disease

| Multiclass Radiomics Disease Severity Signatures | | | | | |
|---|---|---|---|---|---|
| **Features** | **A. Mild (n = 12)** | **H. Moderate (n = 43)** | **H. Severe (n = 19)** | **I. Critical (n = 44)** | ***P*** |
| Laboratory tests | | | | | |
| Lymphocytes | 0.88 ± 0.66 | 0.97 ± 0.46 | 1.09 ± 0.45 | 1.57 ± 1.03 | <0.001 |
| Neutrophils | 5.11 ± 2.63 | 5.33 ± 3.41 | 6.94 ± 3.54 | 9.75 ± 5.02 | 0.015 |
| WBCsn | 7.36 ± 2.02 | 6.73 ± 3.82 | 8.65 ± 4.15 | 12.13 ± 5.62 | 0.0012 |
| eGFR | 76.16 ± 24.15 | 83.50 ± 16.33 | 81.58 ± 16.91 | 53.28 ± 29.4 | <0.001 |
| CT-based quantification of lung lesions | | | | | |
| GGO lesion extent | 6.99 ± 9.31 | 22.28 ± 20.50 | 34.85 ± 19.52 | 31.62 ± 21.12 | 0.039 |
| CON lesion extent | 0.90 ± 1.48 | 6.56 ± 10.4 | 20.61 ± 22.3 | 27.65 ± 24.08 | <0.001 |
| PLE lesion extent | 0.00 ± 0.00 | 1.54 ± 4.59 | 0.11 ± 0.37 | 5.07 ± 9.29 | <0.001 |
| BAN lesion extent | 0.25 ± 0.34 | 0.45 ± 0.87 | 0.17 ± 0.46 | 0.16 ± 0.39 | <0.001 |
| Radiomic markers | | | | | |
| GGO lesion extent | −0.83 (−0.92, −0.74) | −0.40 (−0.48, −0.33) | 0.15 (0.02, 0.27) | 0.53 (0.44, 0.63) | <0.001 |
| GGO F.O. maximum | −0.55 (−0.74, −0.37) | −0.14 (−0.23, −0.04) | 0.31 (0.18, 0.45) | 0.16 (0.07, 0.24) | <0.001 |
| GGO major axis length | −0.57 (−0.80, −0.35) | −0.08 (−0.18, 0.02) | 0.10 (−0.01, 0.23) | 0.13 (0.04, 0.22) | <0.001 |
| GGO difference variance | −0.58 (−0.78, −0.38) | −0.02 (−0.13, 0.08) | 0.19 (0.09, 0.30) | −0.04 (−0.12, 0.04) | <0.001 |
| GGO joint average | −0.68 (−0.88, −0.48) | −0.18 (−0.28, −0.08) | 0.32 (0.20, 0.43) | 0.15 (0.06, 0.24) | <0.001 |
| GGO size zone non-UI | −0.74 (−0.83, −0.65) | −0.36 (−0.43, −0.29) | 0.21 (0.06, 0.37) | 0.42 (0.32, 0.52) | <0.001 |
| CON lesion extent | −0.50 (−0.55, −0.44) | −0.32 (−0.36, −0.28) | 0.01 (−0.09, 0.11) | 0.36 (0.23, 0.49) | <0.001 |
| CON F.O. kurtosis | −0.65 (−0.75, −0.54) | −0.19 (−0.27, −0.11) | −0.08 (−0.16, −0.01) | 0.40 (0.28, 0.51) | <0.001 |
| CON cluster tendency | −0.29 (−0.37, −0.21) | 0.01 (−0.11, 0.12) | 0.02 (−0.04, 0.08) | 0.10 (−0.01, 0.21) | 0.003 |
| CON cluster tendency | −0.29 (−0.37, −0.21) | 0.01 (−0.11, 0.12) | 0.02 (−0.04, 0.08) | 0.10 (−0.01, 0.21) | 0.003 |
| CON joint average | −0.89 (−1.0, −0.78) | −0.21 (−0.30, −0.11) | 0.03 (−0.09, 0.15) | 0.40 (0.31, 0.48) | <0.001 |
| CON zone entropy | −1.08 (−1.21, −0.87) | −0.23 (−0.33, −0.13) | 0.14 (0.017, 0.27) | 0.45 (0.38, 0.52) | <0.001 |
| PLE lesion extent | −0.30 (−0.31, −0.29) | −0.08 (−0.17, 0.01) | −0.30 (−0.32, −0.28) | 0.30 (0.18, 0.43) | <0.001 |
| PLE F.O. 90th percentile | −0.33 (−0.35, −0.32) | −0.17 (−0.27, −0.06) | −0.22 (−0.32, −0.12) | 0.34 (0.24, 0.44) | <0.001 |
| PLE busyness | −0.28 (−0.29, −0.27) | −0.10 (−0.18, −0.03) | −0.29 (−0.30, −0.28) | 0.32 (0.18, 0.45) | <0.001 |
| PLE dependence entropy | −0.48 (−0.512, −0.46) | −0.20 (−0.28, −0.12) | −0.35 (−0.45, −0.26) | 0.43 (0.32, 0.54) | <0.001 |
| BAN lesion extent | 0.01 (−0.10, 0.13) | 0.01 (−0.03, 0.23) | 0.02 (−0.14, 0.18) | −0.12 (−0.14, −0.1) | 0.013 |
| BAN difference entropy | −0.017 (−0.19, 0.15) | 0.034 (−0.06, 0.13) | −0.05 (−0.18, 0.08) | −0.10 (−0.19, −0.01) | 0.021 |

The link between laboratory test results, CT-based quantification of lung lesions, and radiomics signatures in characterizing the 4 proposed stages of COVID-19 severity.

Note: Disease extent and laboratory variables are represented as median and standard deviation. The laboratory test results were obtained within ±12 hours of CT acquisition per subject. Radiomics variables are represented as the median and interquartile range (IQR). Radiomics values fall within a range of −1 to 1. The *P* values were obtained through an analysis of variance (ANOVA) calculation.

WBCsn, white blood cell count; eGFR, estimated glomerular filtration rate; CT, computed tomography; GGO, ground-glass opacity; CON, consolidation; PLE, pleural effusion; BAN, band-like structure.

had a greater extent of GGO than those in the critical intubated condition. This could be explained by a higher presence of CON and PLE lesions in critically intubated patients and the inclusion of intubated patients with organ failure (as indicated by a WHO-9 score) in the study population.

Figure S5 in the Supplementary Material, http://links.lww.com/RLI/A833, compares the disease extent estimated by radiologists with that generated by the multiclass model (AssessNet-19). Results show excellent agreement for overall disease extent, GGO, and CON, with Cohen κ values of 0.94/0.87, 0.90/0.71, and 0.83/0.71, respectively. The multiclass model combines CON with PLE, as the radiologists' qualitative CON disease estimation includes PLE.

## Disease Severity Assessment Benchmarking

In assessing COVID-19 disease severity, the multiclass model proved superior to the single-class model and the radiologists' qualitative score. Using 4-severity states, the multiclass model achieved an F1-score of 0.76 ± 0.02, whereas the radiologists and single-class model scored 0.63 ± 0.02 and 0.64 ± 0.02, respectively. Figure 6A



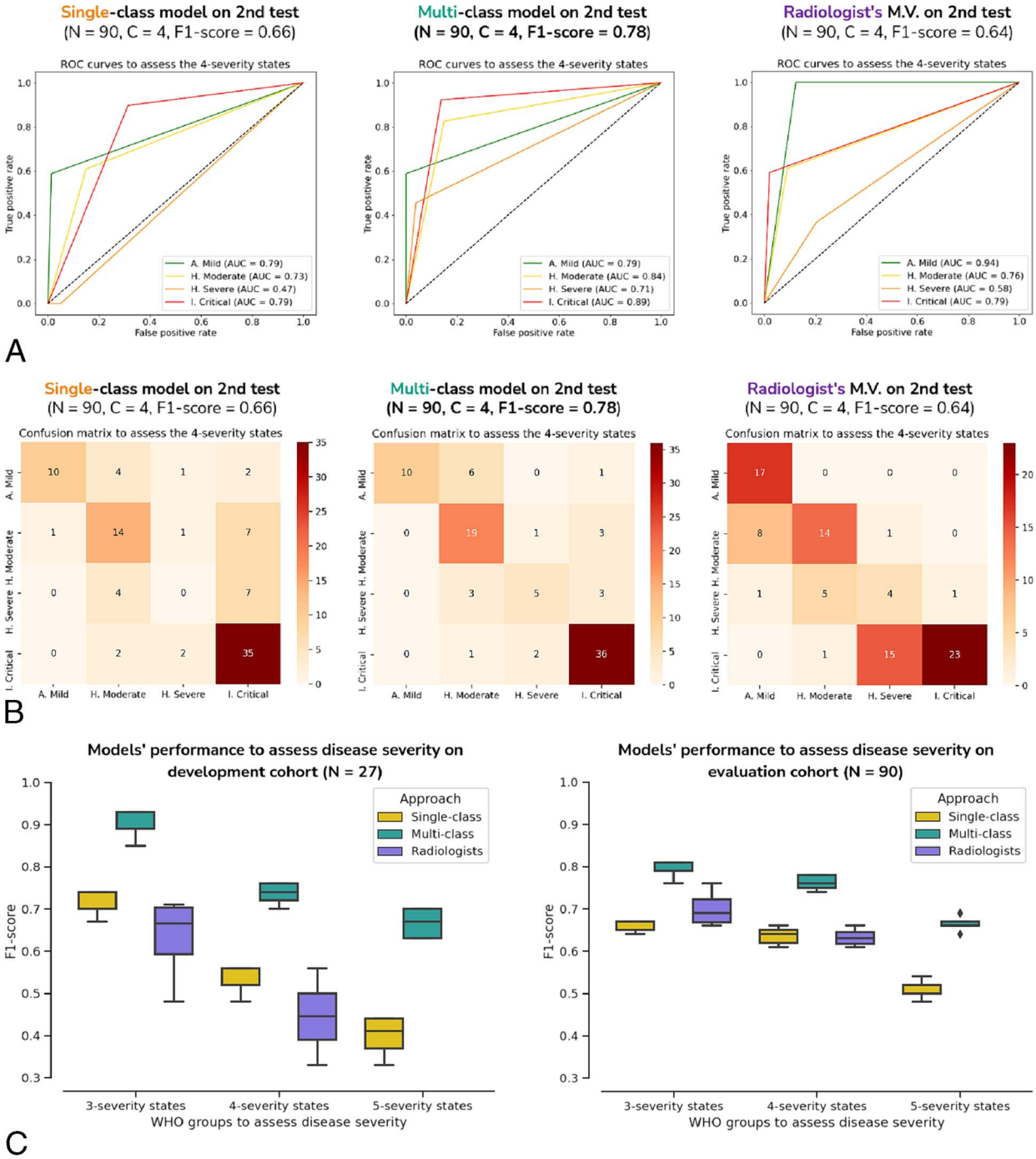


**FIGURE 6.** Comparison of a single-class model versus a multiclass model and the qualitative score from 3 expert radiologists to assess the disease state of COVID-19 patients.

displays the patients' evaluation outcomes for the 4 disease severity levels in the evaluation set.

Furthermore, we used ROC curves to classify each severity level and compared the single-class and multiclass models and the radiologists' subjective scoring. For the ambulatory mild disease state, the radiologists achieved excellent classification performance with an AUC of 0.94, whereas the multiclass and single-class models had good performance with AUCs of 0.79 and 0.79, respectively. The multiclass model showed excellent classification performance for the HM state with an AUC of 0.84, whereas the radiologists and single-class model achieved good performance with AUCs of 0.76 and 0.73, respectively. For the HS state, the multiclass model achieved good performance with an AUC of 0.71, the radiologists achieved OK performance with an AUC of 0.58, and the single-class model achieved poor performance with an AUC of 0.47. The multiclass model also achieved excellent classification performance for the IC state with an AUC of 0.89, whereas the radiologists and single-class model had good performance with 0.79 and 0.79, respectively. Finally, we present the corresponding confusion matrices for the single-class and multiclass lesion models and the radiologists' qualitative scores on the evaluation cohort in Figure 6B.



### Comparing Hierarchical Multilabel Classification for Grouping WHO Severity Scores

In this study, we collapsed the WHO scale into 3, 4, and 5 hierarchical labels to evaluate the performance of the multiclass and single-class model. These results are presented in box plots in Figure 6C and Table 3, demonstrating the performance of disease severity assessment across all severity states. The box plots visually represent the variation in performance of F1 score for each severity state. For the 3-label hierarchy, the multiclass model achieved an F1-score of 0.77, correctly classifying 12 of 18 MA cases, 30 of 35 HD cases, and 30 of 39 IC cases. In contrast, the single-class model achieved an F1-score of 0.65, correctly classifying 8 of 18 MA cases, 22 of 35 HD cases, and 30 of 39 IC cases. In the 4-label hierarchy, the multiclass model achieved an F1-score of 0.76, correctly classifying 10 of 18 MA cases, 19 of 24 HM cases, 5 of 11 HS cases, and 36 of 39 IC cases. The single-class model achieved an F1-score of 0.64, correctly classifying 10 of 18 MA cases, 14 of 24 HM cases, 0 of 11 HS cases, and 35 of 39 IC cases. Lastly, in the 5-label hierarchy, the multiclass model achieved an F1-score of 0.65, correctly classifying 11 of 18 MA cases, 19 of 24 HM cases, 5 of 11 HS cases, 22 of 32 IC cases, and 4 of 7 IC+ cases. The single-class model achieved an F1-score of 0.51, correctly classifying 10 of 18 MA cases, 14 of 24 HM cases, 1 of 11 HS cases, 21 of 32 IC cases, and 1 of 7 IC+ cases. These results demonstrate the performance of both models across different hierarchical label configurations for classifying the severity of COVID-19 patients based on the WHO scores. Please see Section S2 in the Supplementary Material, http://links.lww.com/RLI/A833, for the confusion matrices that compare hierarchical multilabel classification between the single-class and multiclass models. For a detailed and comprehensive comparison, please refer to Figure 3 in the Supplementary Material, http://links.lww.com/RLI/A833.

## DISCUSSION

This study presents AssessNet-19, a multiclass radiomics model that accurately assesses COVID-19 severity. Compared with traditional models and radiologists' evaluations, the model achieved an F1-score of $0.76 \pm 0.02$, which is 12% higher than the single-class model's F1-score and 13% higher than the radiologists' majority vote F1-score. The model uses the WHO-CPS to classify the severity of COVID-19 cases and separates lung lesions into 4 categories: GGO, CON, PLE, and BAN. The 2D-UNet model was trained using a sparse annotation strategy to improve efficiency and reduce annotation time. AssessNet-19 automates CT image segmentation to produce a severity score and radiomics signature for characterizing a patient's status. This study establishes a standardized, automated workflow for classifying COVID-19 cases based on WHO guidelines, enabling future disease characterization and prediction research.

The WHO introduced the WHO-CPS for COVID-19 in 2020 as a standard evaluation basis for cohort studies and clinical trials and to aid in resource planning.[16] Although this scale has been used in several studies, to our knowledge, no study has applied AI and chest imaging to assess the severity proposed by the WHO. Ramaswamy et al[26] detailed the calculation of the WHO score from electronic medical records but did not use AI techniques or imaging. Bennett et al[27] used the WHO scale and grouped severity categories to predict clinical severity and found that demographics and comorbidities were correlated with disease severity. Our study compares the performance of a single-class and multiclass segmentation model in classifying the severity of COVID-19 patients using the WHO-CPS. It found that the multiclass model had better results, with a 10%, 11%, and 15% greater F1-score than the single-class model, when the WHO-CPS was grouped into 3, 4, and 5 classes, respectively.

In this study, the multiclass AI-model achieved a higher F1-score than radiologists in assessing COVID-19 patients using CT imaging. This is consistent with previous research that shows AI models perform similarly or better than radiologists in assessing lung diseases.[28–30] In addition, multiple studies have demonstrated that AI models using medical imaging are more accurate in predicting the progression or outcome of severe clinical states than radiologists' scores.[30–32] In this study, the multiclass AI model outperformed radiologists in assessing disease severity using both 3-class and 4-class WHO scores (radiologists did not evaluate a 5-class WHO score), particularly in differentiating between hospitalized severe and critical intubated cases. This could be due to radiologists evaluating images based solely on imaging findings without regular feedback regarding the clinical status of those patients, which tends to underestimate the clinical severity represented by the pathology seen in the images. The AI model, on the other hand, learns to evaluate imaging characteristics based on their consequences for patients' clinical condition rather than the extent of pathology, which is a significant benefit of using a supervised learning approach with the WHO-CPS outcomes to train a medical imaging AI.

Moreover, this study demonstrated that using a 2D-UNet model called AssessNet-19, along with a sparse annotation strategy, improves computational efficiency and reduces expert annotation time when compared with a 3D-UNet model. Furthermore, the study conducted single- and multiclass lesion segmentation using the AssessNet-19 model and compared it with 2 other state-of-the-art models: a 3D single lesion model based on a full-resolution 3D U-Net trained with the COVID-19 2020 grand challenge data set and a Scancovia segmentation model presented by Lassau et al.[14,33] The results of the study revealed that AssessNet-19 outperformed RapidSegLesion-19 in single lesion segmentation, achieving a higher Dice score of 0.77 compared with RapidSegLesion-19's Dice score of 0.65. AssessNet-19 showed particular effectiveness in accurately segmenting lesions in dense areas. In addition, the study compared AssessNet-19 with Scancovia for multiclass lesion segmentation. AssessNet-19 demonstrated better performance in segmenting COVID-19 lesions, with higher Dice scores of 29%, 24%, and 50% for single, GGO, and CON lesions, respectively. On the other hand, Scancovia exhibited issues with oversegmentation in mild cases and undersegmentation in severe cases, similar to RapidSegLesion-19. It is important to note that caution should be exercised when interpreting this comparison since the 2 models were trained using different segmentation protocols. For more detailed information, please refer to Section S3 of the Supplementary Material, http://links.lww.com/RLI/A833.

The multiclass lesion segmentation protocol proposed in the study, which classifies lung lesions into 4 categories (GGO, CON, PLE, and BAN), allows the AssessNet-19 model to be more accurate than radiologists' evaluations or traditional models. TBR segmentations

were included as an additional class in the segmentation model but were yet to be used in the severity assessment phase of AssessNet-19 because quantification of bronchi and differentiation between enlarged and normal bronchi required more work and will be performed in the future. The study found significant differences in the classification of the 4 disease states for BAN ($P < 0.001$), with disease severity increasing from mild to severe and decreasing for critical cases. This observation could be attributed to the extent of CONs in critical cases increases, obscuring the BANs. A recent study shows the significance of BAN, with pleuroparenchymal bands present in 36 of 42 long COVID patients 3 months postacute phase.[34] However, more research is needed to understand the role of BAN in long COVID and disease progression.

Previous research has shown that deep learning on medical imaging can accurately predict COVID-19 outcomes and assess disease severity. Some studies focus on using DL models for quantifying CT patterns to identify abnormalities in COVID-19 patients, whereas others aim to improve results by combining clinical data and images to identify severe outcomes.[10,12–15,31,35] In addition, we automated the multiclass radiomics model AssessNet-19 to address CT quantification, severity assessment, and disease characterization. On the other hand, other studies uses radiomics features extracted from medical images to predict disease outcomes and used the radiomics features to characterize COVID-19 severity.[36–38] We advanced prior studies by categorizing COVID-19 severity into 4-disease statuses using the WHO-CPS. We also characterized COVID-19 lesions in 4-lung pathologies and created a "severity signature" by utilizing radiomics characteristics from each lung pathology. This enabled the interpretation and quantification of patient status, visualization of disease-population values, and analysis of lesion and interlesion patterns in radiological findings.

We selected LASSO for feature selection and XGBoost for classification based on the state of the art, careful consideration, and empirical evaluation. LASSO was chosen for feature selection due to its effective handling of high-dimensional data and ability to identify relevant features. It performs feature selection and regularization, reducing overfitting and improving interpretability. Furthermore, we conducted a radiomics normalization comparison to improve classification performance. The experiment compared unnormalized radiomic features with $z$-score normalization, and the 2-step normalization approach exhibited superior performance, consistent with the findings reported in.[39] For classification, we compared XGBoost and Random Forest classifiers. Our preliminary experiments showed that XGBoost outperformed Random Forest, achieving an F1-score of 0.75 compared with 0.59. XGBoost's strength in handling complex relationships, missing data, and capturing nonlinearities made it a suitable choice. We also evaluated voting approaches (majority and average voting) for ranking the severity label. Both classifiers performed better with majority voting, indicating its superiority in capturing consensus. In summary, we chose LASSO for feature selection and XGBoost for classification due to their proven effectiveness in handling high-dimensional data, selecting informative features, mitigating overfitting, capturing complex relationships, and achieving superior classification performance. For a detailed comparison of feature importance between the single and multiclass models, including SHapley Additive exPlanations (SHAP values), please refer to Figure S4 (Supplementary Material, http://links.lww.com/RLI/A833).

The study has limitations because of a small sample size. There were also a few ambulatory-mild cases (WHO-3) in our data set, which may explain the model's poor performance in classifying them. This may be because patients with mild symptoms are less likely to receive a CT. Analysis of misclassified subjects can be found in the Supplementary Material, http://links.lww.com/RLI/A833. Although our model was trained and tested on a quite diverse data set, further validation would have required an external data set despite using a diverse data set. The lack of clinical data in cases transferred from other hospitals affects all longitudinal studies and presents challenges in including patients in our study.

In conclusion, the proposed AssessNet-19 model based on multiclass lesion segmentation combined with WHO-standardized severity scaling and radiological characterization is the first step to developing a generalizable clinical decision support system for hospitals not only during the COVID-19 pandemic but also for similar challenges in the future.

## ACKNOWLEDGMENTS

*The authors thank the support provided by the Swiss National Science Foundation through the National Research Programme "COVID-19" (NRP 78) under grant number 198388, as well as Campus Stiftung Lindenhof Bern and the Swiss Institute for Translational and Entrepreneurial Medicine for their valuable support throughout this research project.*

## REFERENCES

1. Ceruti S, Glotta A, Biggiogero M, et al. Admission criteria in critically ill COVID-19 patients: a physiology-based approach. *PLoS One*. 2021;16:e0260318.
2. Wan S, Li M, Ye Z, et al. CT manifestations and clinical characteristics of 1115 patients with coronavirus disease 2019 (COVID-19): a systematic review and meta-analysis. *Acad Radiol*. 2020;27:910–921.
3. Lyu P, Liu X, Zhang R, et al. The performance of chest CT in evaluating the clinical severity of COVID-19 pneumonia: identifying critical cases based on CT characteristics. *Invest Radiol*. 2020;55:412–421.
4. Born J, Beymer D, Rajan D, et al. On the role of artificial intelligence in medical imaging of COVID-19. *Patterns (N Y)*. 2021;2:100269.
5. Baghdadi NA, Malki A, Abdelaliem SF, et al. An automated diagnosis and classification of COVID-19 from chest CT images using a transfer learning-based convolutional neural network. *Comput Biol Med*. 2022;144:105383.
6. Arora R, Bansal V, Buckchash H, et al. AI-based diagnosis of COVID-19 patients using x-ray scans with stochastic ensemble of CNNs. *Phys Eng Sci Med*. 2021;44:1257–1271.
7. Pennisi M, Kavasidis I, Spampinato C, et al. An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. *Artif Intell Med*. 2021;118:102114.
8. Lanza E, Muglia R, Bolengo I, et al. Quantitative chest CT analysis in COVID-19 to predict the need for oxygenation support and intubation. *Eur Radiol*. 2020;30:6770–6778.
9. Diniz JOB, Quintanilha DBP, Santos Neto AC, et al. Segmentation and quantification of COVID-19 infections in CT using pulmonary vessels extraction and deep learning. *Multimed Tools Appl*. 2021;80:29367–29399.
10. Chaganti S, Grenier P, Balachandran A, et al. Automated quantification of CT patterns associated with COVID-19 from chest CT. *Radiol Artif Intell*. 2020;2:e200048.
11. Fontanellaz M, Ebner L, Huber A, et al. A deep-learning diagnostic support system for the detection of COVID-19 using chest radiographs: a multireader validation study. *Invest Radiol*. 2021;56:348–356.
12. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. 2022. Available at: https://arxiv.org/abs/2206.01653. Cited January 12, 2023.
13. Chassagnon G, Vakalopoulou M, Battistella E, et al. AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. *Med Image Anal*. 2021;67:101860.
14. Lassau N, Ammari S, Chouzenoux E, et al. Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nat Commun*. 2021;12:634.
15. Soda P, D'Amico NC, Tessadori J, et al. AIforCOVID: predicting the clinical outcomes in patients with COVID-19 applying AI to chest-x-rays. An Italian multicentre study. *Med Image Anal*. 2021;74:102216.
16. Marshall JC, Murthy S, Diaz J, et al. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect Dis*. 2020;20:e192–e197.
17. Hofmanninger J, Prayer F, Pan J, et al. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. 2020. Available at: https://arxiv.org/abs/2001.11767. Cited January 12, 2023.
18. Hansell DM, Bankier AA, MacMahon H, et al. Fleischner society: glossary of terms for thoracic imaging. *Radiology*. 2008;246:697–722.
19. Strickland E. Andrew Ng, AI minimalist: the machine-learning Pioneer says small is the new big. *IEEE Spectr*. 2022;59:22–50.
20. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18:203–211.
21. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, et al, eds. *Medical Image*

*Computing and Computer-Assisted Intervention—MICCAI 2015 [Internet]*. Cham, Switzerland: Springer International Publishing; 2015:234–241. Available at: http://link.springer.com/10.1007/978-3-319-24574-4_28. Cited January 12, 2023.

22. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77:e104–e107.
23. Muthukrishnan R, Rohini R. LASSO: a feature selection technique in predictive modeling for machine learning. In: *2016 IEEE International Conference on Advances in Computer Applications (ICACA) [Internet]*. Coimbatore, India: IEEE; 2016:18–20. Available at: http://ieeexplore.ieee.org/document/7887916/. Cited January 12, 2023.
24. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295:328–338.
25. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]*. San Francisco, CA: ACM; 2016:785–794. Available at: https://dl.acm.org/doi/10.1145/2939672.2939785. Cited January 12, 2023.
26. Ramaswamy P, Gong JJ, Saleh SN, et al. Developing a COVID-19 WHO Clinical Progression Scale inpatient database from electronic health record data. *J Am Med Inform Assoc*. 2022;29:1279–1285.
27. Bennett TD, Moffitt RA, Hajagos JG, et al. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US national COVID cohort collaborative. *JAMA Netw Open*. 2021;4:e2116901.
28. Li MD, Arun NT, Gidwani M, et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional Siamese neural networks. *Radiol Artif Intell*. 2020;2:e200079.
29. Mushtaq J, Pennella R, Lavalle S, et al. Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients. *Eur Radiol*. 2021;31:1770–1779.
30. Fang X, Kruger U, Homayounieh F, et al. Association of AI quantified COVID-19 chest CT and patient outcome. *Int J Comput Assist Radiol Surg*. 2021;16:435–445.
31. Jiao Z, Choi JW, Halsey K, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. *Lancet Digit Health*. 2021;3:e286–e294.
32. Homayounieh F, Ebrahimian S, Babaei R, et al. CT Radiomics, radiologists, and clinical information in predicting outcome of patients with COVID-19 pneumonia. *Radiol Cardiothorac Imaging*. 2020;2:e200322.
33. Roth H, Xu Z, Diez CT, et al. Rapid artificial intelligence solutions in a pandemic—the COVID-19-20 lung CT lesion segmentation challenge [Internet]. *In Review*. 2021. Available at: https://www.researchsquare.com/article/rs-571332/v1. Cited January 12, 2023.
34. Bocchino M, Lieto R, Romano F, et al. Chest CT–based assessment of 1-year outcomes after moderate COVID-19 pneumonia. *Radiology*. 2022;305:479–485.
35. Mei X, Lee HC, Diao KY, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat Med*. 2020;26:1224–1228.
36. Bermejo-Peláez D, San José Estépar R, Fernández-Velilla M, et al. Deep learning-based lesion subtyping and prediction of clinical outcomes in COVID-19 pneumonia using chest CT. *Sci Rep*. 2022;12:9387.
37. Bouchareb Y, Moradi Khaniabadi P, Al Kindi F, et al. Artificial intelligence-driven assessment of radiological images for COVID-19. *Comput Biol Med*. 2021;136:104665.
38. Li W, Cao Y, Yu K, et al. Pulmonary lesion subtypes recognition of COVID-19 from radiomics data with three-dimensional texture characterization in computed tomography images. *Biomed Eng Online*. 2021;20:123.
39. Chen J, Cheung HMC, Milot L, et al. AMINN: autoencoder-based multiple instance neural network improves outcome prediction in multifocal liver metastases. In: de Bruijne M, Cattin PC, Cotin S, et al Medical Image Computing and Computer Assisted Intervention—MICCAI 2021. Lecture Notes in Computer Science. Cham, Switzerland: Springer; 2021:12905.