High-dimensional multinomial multiclass severity scoring of COVID-19 pneumonia using CT radiomics features and machine learning algorithms

Isaac Shiri1, Shayan Mostafaei2, Atlas Haddadi Avval3, Yazdan Salimi1, Amirhossein Sanaat1, Azadeh Akhavanallaf1, Hossein Arabi1, Arman Rahmim4,5 & Habib Zaidi1,6,7,8*

We aimed to construct a prediction model based on computed tomography (CT) radiomics features to classify COVID-19 patients into severe-, moderate-, mild-, and non-pneumonic. A total of 1110 patients were studied from a publicly available dataset with 4-class severity scoring performed by a radiologist (based on CT images and clinical features). The entire lungs were segmented and followed by resizing, bin discretization and radiomic features extraction. We utilized two feature selection algorithms, namely bagging random forest (BRF) and multivariate adaptive regression splines (MARS), each coupled to a classifier, namely multinomial logistic regression (MLR), to construct multiclass classification models. The dataset was divided into 50% (555 samples), 20% (223 samples), and 30% (332 samples) for training, validation, and untouched test datasets, respectively. Subsequently, nested cross-validation was performed on train/validation to select the features and tune the models. All predictive power indices were reported based on the testing set. The performance of multi-class models was assessed using precision, recall, F1-score, and accuracy based on the 4 × 4 confusion matrices. In addition, the areas under the receiver operating characteristic curves (AUCs) for multi-class classifications were calculated and compared for both models. Using BRF, 23 radiomic features were selected, 11 from first-order, 9 from GLCM, 1 from GLRM, and 1 from shape. Ten features were selected using the MARS algorithm, namely 3 from first-order, 1 from GLDM, 1 from GLRLM, 1 from GLSZM, 1 from shape, and 3 from GLCM features. The mean absolute deviation, skewness, and variance from first-order and flatness from shape, and cluster prominence from GLCM features and Gray Level Non Uniformity Normalize from GLRLM were selected by both BRF and MARS algorithms. All selected features by BRF or MARS were significantly associated with four-class outcomes as assessed within MLR (All p values < 0.05). BRF + MLR and MARS + MLR resulted in pseudo-R² prediction performances of 0.305 and 0.253, respectively. Meanwhile, there was a significant difference between the feature selection models when using a likelihood ratio test (p value = 0.046). Based on confusion matrices for BRF + MLR and MARS + MLR algorithms, the precision was 0.856 and 0.728, the recall was 0.852 and 0.722, whereas the accuracy was 0.921 and 0.861, respectively. AUCs (95% CI) for multi-class classification were 0.846 (0.805–0.887) and 0.807 (0.752–0.861) for BRF + MLR and MARS + MLR algorithms, respectively. Our models based on the utilization of radiomic features, coupled with machine learning were able to accurately classify patients according
to the severity of pneumonia, thus highlighting the potential of this emerging paradigm in the prognostication and management of COVID-19 patients.

Abbreviations
CT       Computed tomography
COVID-19  Coronavirus disease 2019
AUC      Area under the receiver operating characteristic curve
CV       Cross-validation
BRF      Bagging random forest
FS       Feature selection
IV       Importance value
ML       Machine learning
GGO      Ground glass opacity
IBSI     The image biomarker standardization initiative
MARS     Multivariate adaptive regression splines
MLR      Multinomial logistic regression
RT-PCR   Reverse transcription polymerase chain reaction
GLCM     Gray-level co-occurrence matrix
GLSZM    Gray-level size-zone matrix
NGTDM    Neighbouring gray tone difference matrix
GLRLM    Gray-level run-length matrix
GLDM     Gray-level dependence matrix

The highly contagious SARS-CoV-2 virus has led to significant morbidity and mortality worldwide. Pneumonia is regarded as one of the main complications of COVID-19 disease, which can lead to lethal conditions while escalating the cost of healthcare. The most popular diagnostic test considered as the gold standard for coronavirus disease is the reverse transcription polymerase chain reaction (RT-PCR) assay. While highly specific, RT-PCR has shown low sensitivity, as studies have reported significant false-negatives in patients who had abnormalities in their chest CT images confirmed with secondary follow-up RT-PCR to be positive for COVID-19.

CT aids in the diagnosis and management of COVID-19 patients and could be potentially used as an outcome/survival prediction tool, towards enhanced treatment planning. CT scanning has been utilized as a highly sensitive tool for COVID-19 diagnosis since it is fast and generates quantifiable features (e.g., the extent to which lung lobes are involved) and non-quantifiable features (e.g., ground-glass opacities and their laterality) to assess COVID-19 pneumonia, besides the enhanced sensitivity compared to RT-PCR.

Severity can be defined as an index that depicts the effects of a disease on mortality, morbidity, and comorbidities and has the potential to help physicians manage the patients more decently whether in patients with cancer or with non-cancer diseases. A number of severity scoring systems have been proposed to quantify disease advancement in patients, including general assessments (e.g., APACHE score) and disease-specific ones (e.g., Child–Pugh score). Several conventional scoring systems have been proposed for COVID-19 severity assessment. These include the usage of patient clinical, comorbidity, and laboratory data, which are all helpful in constructing predictive models for severity assessment in COVID-19.

There has also been a growing interest in using imaging data of patients, such as thoracic CT images. For example, a study by Sanders et al. computed the score of CT images in patients with cystic fibrosis and evaluated the prognostic ability. A promising line of research that emerged recently reported on the CT severity index and its correlation with acute pancreatitis severity. The COVID-19 Reporting and Data System (CO-RADS) was suggested for standardized visual assessment of COVID-19 pneumonia to enhance agreement between radiologists. This system includes features for the diagnosis of COVID-19 and consists of a 6-point scale for categorizing patient CT images. In addition, other guidelines aiming to reach consensus when interpreting COVID-19 suspected chest CT images were proposed. These guidelines are mostly based on visual assessment of images; e.g., the amount to which lung lobes are involved, the volume of which is infected, and anatomical assessments.

Francone et al. reported a study on the correlation between CT score and the severity of coronavirus disease. Zhao et al. also conducted research on the measurement of the extent to which lung lobes are infected and evaluation in COVID-19 patients’ prognosis. Li et al. also confirmed the association between chest CT score and COVID-19 pneumonia severity. At the same time, most scoring systems involve visual assessment and hence are time-consuming. In this regard, medical image analysis using machine learning (ML) and radiomics has been applied to quantify features to tackle these main challenges.

The field of radiomics opens pathways for the study of normal tissues, cancer, cardiac disease, and many other diseases, including potentially the newly emerging COVID-19 disease. Specifically, Xie et al. evaluated the potential of a radiomics framework to diagnose COVID-19 from CT images. Di et al. also studied whether radiomics features can help to distinguish between pneumonia of COVID-19 and that of other viral/bacterial causes. A number of studies reported on the application of radiomics analysis to CT images towards COVID-19 classification and prognostication. Homayounieh et al. assessed the prognostic power of CT-based radiomics features to determine severe and non-severe cases. In another study, Li et al. proposed a radiomics model based on CT images and classified patients based on the criticality of their disease. A recent study by Yip et al. applied a robust radiomics model to CT images to predict the severity of COVID-19 disease in patients. All above models pursued binary task performance, which reduced multiclass classification to two class approaches.
However, in the real clinical triage situation, scoring systems consist of multi-class datasets. In the present study, involving a large cohort of patients, we aimed to construct a CT radiomics-based multi-class classification model to predict the severity of COVID-19 pneumonia.

**Materials and methods**

**Data description.** Figure 1 presents the different steps performed in this study. All experiments were performed in accordance with relevant guidelines and regulations.

**Datasets and segmentation.** This study is based on the MosMed Dataset consisting of 1110 patient CT scans, also utilized in other efforts. Ethics approval and consent to participate were not needed since the study was performed on open access online dataset. In the class zero, the patient has neither clinical symptoms (e.g. fever) nor CT findings in favor of any kind of pneumonia (Class 0, non-pneumonic). The 1st class contains patients who have a low-temperature fever (t < 38 °C) in addition to a mild increase in respiratory rate (RR < 20) while showing none or < 25% ground-glass opacity (GGO) involvement (Class 1, COVID-19 with mild severity). Patients in the 2nd class have a higher body temperature (t > 38.5 °C) with a RR of 20–30, while CT scan shows 25–50% involvement of lung parenchyma (Class 2, COVID-19 with moderate severity). Patients in the 3rd class have high body temperature and RR of 30 or more, with CT findings of 50% to diffuse involvement in addition to organ failure and shock signs (Class 3, severe COVID-19). Each of the classes, namely 0, 1, 2, and 3,
included 254, 684, 125, and 47 patients, respectively. The median age was 47 (ranging from 18 to 97), and 42% of patients were female. Figure 2 shows an example of representative CT images for each class.

All CT images were automatically segmented using a deep learning-based algorithm for whole lung segmentation. After whole-lung 3D segmentation, all images were reviewed and modified to ensure correct 3D-volume lung segmentation.

**Image preprocessing and feature extraction.** To preserve image resolution and efficient radiomics feature extraction, all images were cropped to lung region and then resized to 296 × 216 matrix size. Subsequently, image voxels were resized to an isotropic voxel size of 1 × 1 × 1 mm³ (for invariant texture feature extraction) and image intensity were discretized to 64-binning size. The extracted features from the whole-lung segmented regions, totalling 110, included shape (n = 16), intensity (n = 19), and texture features, namely second-order texture of gray-level co-occurrence matrix (GLCM, n = 24), and high-order features, namely gray-level size-zone matrix (GLSZM, n = 16), neighbouring gray tone difference matrix (NGTDM, n = 5), gray-level run-length matrix (GLRLM, n = 16) and gray-level dependence matrix (GLDM, n = 14). Radiomics feature extraction was performed using the Pyradiomics Python library, which is compliant with the image biomarker standardization initiative (IBSI). In addition, feature maps were generated using voxelwise feature extraction.

**Feature selection and classification and evaluation.** In this study, we used two different feature selection algorithms, including Bagging Random Forests (BRF) and Multivariate Adaptive Regression Splines (MARS). BRF and MARS algorithms were implemented in “VSURF” and “earth” R packages, respectively. Importance values (IVs) were calculated using generalized cross-validation criterion with normalization. For multiclass classification, we implemented multinomial logistic regression (MLR) using the “mnlogit” R package. The MLR model fitness indices included p value of the Wald test (corrected for false-discovery rate via Benjamin and Hochberg method), pseudo R², as well as Akaike information criterion (AIC, goodness of fit indices in generalized linear regression models). In the MLR model, class 0 served as a reference class whereas statistical comparison between two predictive models was performed by the Likelihood Ratio Test.

The dataset was divided into 50% (555 samples), 20% (223 samples), and 30% (332 samples) as training, validation, and untouched test datasets, respectively. The nested fivefold cross-validation with grid search was used to validate models and estimate tuning hyper-parameters based on the minimization of GCV error rate. In our nested fivefold cross-validation processing, there were 5 outer folds (i.e., training and testing sets) and 5 inner folds (i.e., training and validation sets) where the total number of trained models was 25 for each classifier. We report mean precision, recall, F1-score, and accuracy and their standard deviation (SD) for different classes in each model based on the 30% untouched test set with bootstrapping (n = 1000) to ensure reproducibility. In addition, the areas under the receiver operating characteristic (ROC) curve (AUCs) for multi-class classification models were calculated and compared for both models using “multiROC” and “pROC” R packages, respectively.
Results

Table 1 summarizes the selected features and their relative importance value by BRF and MARS for multiclass classification. These features were selected in train/validation sets using nested cross validation and grid searches. Twenty-three radiomic features were selected by BRF, including 11 from first-order, 9 from GLCM, 1 from GLRLM, 1 from GLDM, and 1 from shape features. Among these features, Correlation (IV: 100%) and Cluster Tendency (IV: 88%) from GLCM, Mean Absolute Deviation (IV: 80%), Robust Mean Absolute Deviation (IV: 72%) and variance (IV: 70%) from first-order features were selected as the most important ones. In the MARS algorithm, 10 features were selected with high IVs, including 2 from first-order, 1 from GLDM, and 1 from GLCM. The highest IV was achieved by Mean Absolute Deviation (IV: 100%) and skewness (IV: 55%) from first-order, Gray Level Variance from GLDM (IV: 53%), and Correlation from GLCM (IV: 54%). The mean absolute deviation, skewness, variance from first-order, flatness from shape, cluster prominence from GLCM features, and Gray Level Non Uniformity Normalize from GLRLM features were selected by both BRF and MARS algorithms. Figure 3 depicts the feature map of different radiomics features in different classes (10Precentile from first order, Gray Level Non-Uniformity Normalized from GLRLM, Idm from GLCM and Zone Entropy from GLSZM).

Table 2 summarizes the adjusted p value (by Benjamini and Hochberg method) of the Wald test and AIC for both feature selection algorithms using MLR model. All selected features yielded a significant p value (< 0.05). BRF + MLR and MARS + MLR resulted in pseudo $R^2$ values of 0.305 and 0.253, respectively. However, there were significant differences between both predictive models when using a likelihood ratio test ($p$ value = 0.046).

Table 3 summarizes classification power indices, including mean (SD) Precision, Recall, F1-score, Accuracy, and AUC via multinomial logistic regression with 1000 bootstrapping samples for each model in untouched test dataset. In terms of F1-score, four-class mean F1-scores were 0.854 and 0.724 for BRF + MLR and MARS + MLR.

| Algorithm                      | Selected variables | Feature type                        | Relative importance value (%) |
|-------------------------------|--------------------|-------------------------------------|-------------------------------|
| **Bagging Random Forests**    | First Order        | Mean Absolute Deviation             | 80                            |
|                               | First Order        | Robust Mean Absolute Deviation      | 72                            |
|                               | First Order        | Variance                            | 70                            |
|                               | First Order        | Interquartile Range                 | 68                            |
|                               | First Order        | Kurtosis                            | 62                            |
|                               | First Order        | Skewness                            | 61                            |
|                               | First Order        | Entropy                             | 42                            |
|                               | First Order        | 10Percentile                       | 40                            |
|                               | First Order        | 90Percentile                       | 36                            |
|                               | First Order        | Energy                              | 30                            |
|                               | First Order        | Mean                                | 20                            |
|                               | GLCM               | Correlation                         | 100                           |
|                               | GLCM               | Cluster Tendency                    | 88                            |
|                               | GLCM               | Sum Squares                         | 66                            |
|                               | GLCM               | Inverse Variance                    | 60                            |
|                               | GLCM               | Cluster Shade                       | 55                            |
|                               | GLCM               | Cluster Prominence                  | 54                            |
|                               | GLCM               | Joint Entropy                       | 52                            |
|                               | GLCM               | Idm                                 | 48                            |
|                               | GLCM               | Id                                  | 44                            |
|                               | GLDM               | Dependence Variance                 | 65                            |
|                               | GLRLM              | Gray Level Non Uniformity Normalize | 51                            |
|                               | Shape              | Flatness                            | 18                            |
| **Multivariate Adaptive Regression Splines** | First Order        | Mean Absolute Deviation             | 100                           |
|                               | First Order        | Skewness                            | 55                            |
|                               | First Order        | Variance                            | 11                            |
|                               | GLCM               | Correlation                         | 54                            |
|                               | GLCM               | Cluster Prominence                  | 47                            |
|                               | GLCM               | Difference Entropy                  | 36                            |
|                               | GLDM               | Gray Level Variance                 | 53                            |
|                               | GLRLM              | Gray Level Non Uniformity Normalize | 10                            |
|                               | GLSZM              | Zone Entropy                        | 20                            |
|                               | Shape              | Flatness                            | 48                            |

Table 1. Selected features by Bagging Random Forests (“VSURF” R package) and multivariate adaptive regression splines (“earth” R package) for multi-class classification using nested fivefold cross validation based on the training set (50% of the samples, N = 555) and the validation set (20% of the samples, N = 223). Relative importance value calculated using generalized cross-validation (GCV) criterion with normalization.
algorithms, respectively. The mean precision was 0.856 and 0.728, whereas the mean recall was 0.852 and 0.722 for BRF + MLR and MARS + MLR algorithms, respectively. BRF + MLR and MARS + MLR algorithms achieved an accuracy of 0.921 and 0.861, respectively, in four-class classification. AUCs (95% CI) for multi-class classification were 0.846 (0.805–0.887) and 0.807 (0.752–0.861) for BRF + MLR and MARS + MLR algorithms, respectively. According to the results of the classification metrics, the predictive power of the BRF + MLR model is higher than MARS + MLR. Figure 4 depicts the confusion matrices for both predictive models based on the testing set whereas Fig. 5 shows the ROC curves for our four-class classification methods.

Discussion

In the current study, we constructed a CT radiomics-based model to predict the severity of COVID-19 patients in a large cohort of patients. To this end, we extracted radiomics features from whole lung segmentations and selected high-importance features utilizing two different algorithms, namely BRF and MARS. The selected features were then fed to a multinomial logistic regression classifier for multiclass severity scoring. We achieved 0.846 (0.805–0.887) and 0.807 (0.752–0.861) for AUC, and 0.921 and 0.861 for accuracy in BRF- and MARS-selected features, respectively. We used an automatic model to segment chest CT images for two reasons. First, most CT scans performed in the COVID-19 pandemic era are low-dose. In addition, these scans are acquired with a high pitch. Hence, it is difficult for radiologists to find and follow lung fissures to manually detect or segment the anatomical lobes. As such, we used our previously constructed and validated deep learning model to fully segment the entire lung of each patient.33,34,42,43,47

Yip et al.39 conducted a study on the same dataset utilized in this work, aiming to evaluate some radiomics features towards severity class prediction in patients. They included all 1110 patient CT scans and extracted 107 radiomics features. The maximum relevance minimum redundancy (MRMR) and recursive feature elimination (RFE) algorithms were exploited for feature selection and analysis of the selected features using univariate and multivariate approaches using a logistic regression model to classify as accurately as possible. In their study, the patients were categorized into three severity categories, namely mild, moderate, and severe, to perform two-class classification tasks (mild vs. severe and moderate vs. severe) by splitting the data into training (60%) and test (40%) sets. The authors obtained an AUC of 0.65 in differentiating between moderate and severe cases, while their model performed better (AUC = 0.85) in distinguishing mild vs. severe forms of COVID-19 disease. In this work, we reached an overall AUC of 0.846. In our study and the one by Yip et al.,39 feature extractions were performed using Pyradiomics40 as applied to the entire lung. Interestingly, there were some commonly selected features arrived at via feature selection in both studies, including Mean Absolute Deviation, 10Percentile,
Table 2. Multinomial logistic regression for the selected features by “mnlogit” R package and the model’s fitness indices based on the testing set (N = 332). p value by Wald chi-square test, Adj. p value: P value adjusted by Benjamini and Hochberg method, statistical comparison between two models showed non-significant difference by Likelihood Ratio Test: P value = 0.046, AIC: Akaike Information criterion.

| Algorithm                      | Feature type                  | Adj. p value | Pseudo R² | AIC    |
|--------------------------------|-------------------------------|--------------|-----------|--------|
| Bagging Random Forests         | First Order Mean Absolute Deviation | < 0.001     |           |        |
|                                | First Order Robust Mean Absolute Deviation | < 0.001     |           |        |
|                                | First Order Variance          | < 0.001     |           |        |
|                                | First Order Interquartile Range | < 0.001     |           |        |
|                                | First Order Kurtosis          | < 0.001     |           |        |
|                                | First Order Skewness          | < 0.001     |           |        |
|                                | First Order Entropy           | 0.001       |           |        |
|                                | First Order 10Percentile      | 0.002       |           |        |
|                                | First Order 90Percentile      | 0.001       |           |        |
|                                | First Order Energy            | 0.005       |           |        |
|                                | First Order Mean              | 0.025       |           |        |
|                                | GLCM Correlation              | < 0.001     |           |        |
|                                | GLCM Cluster Tendency         | < 0.001     |           |        |
|                                | GLCM Sum Squares              | < 0.001     |           |        |
|                                | GLCM Inverse Variance         | < 0.001     |           |        |
|                                | GLCM Cluster Shade            | < 0.001     |           |        |
|                                | GLCM Cluster Prominence       | < 0.001     |           |        |
|                                | GLCM Joint Entropy            | < 0.001     |           |        |
|                                | GLCM Id                       | 0.001       |           |        |
|                                | GLCM Idm                      | 0.001       |           |        |
|                                | GLDM Dependence Variance      | < 0.001     |           |        |
|                                | GLRLM Gray Level Non-Uniformity Normalize | 0.009      |           |        |
|                                | Shape Flatness                | < 0.001     |           |        |
| Multivariate Adaptive Regression Splines | First Order Mean Absolute Deviation | < 0.001     |           |        |
|                                | First Order Skewness          | < 0.001     |           |        |
|                                | First Order Variance          | < 0.001     |           |        |
|                                | GLCM Correlation              | < 0.001     |           |        |
|                                | GLCM Cluster Prominence       | < 0.001     |           |        |
|                                | GLCM Correlation              | < 0.001     |           |        |
|                                | GLCM Difference Entropy       | < 0.001     |           |        |
|                                | GLDM Gray Level Variance      | < 0.001     |           |        |
|                                | GLRLM Gray Level Non-Uniformity Normalize | 0.001      |           |        |
|                                | Shape Flatness                | < 0.001     |           |        |

Table 3. The classification power indices (SD) based on the testing set (N = 332) with 1000 bootstrapping samples based on the feature selection methods.
90th Percentile, and Mean from first order and Correlation from GLCM. These selected features in both studies could potentially be used as predictors as they provide information about the intensity and heterogeneity of the lung in COVID-19 patients.

A noticeable advantage of the study by Yip et al.\textsuperscript{39} was the use of a second radiologist observer who classified patients’ images into mild, moderate, and severe classes without paying attention to the default classification of the dataset provider. This method helped to observe the prediction power of the models in both “provider” and “radiologist” datasets. In addition, the study by Yip et al.\textsuperscript{39} may have reduced generalizability as it only predicts mild versus severe, and moderate versus severe disease, having reduced multiclass classification into two-class approaches. In the real clinical triage situation, the radiologist may benefit from a multiclass classification scheme for enhanced patient management, as provided by our study.

Multi-class classification is a difficult machine learning task\textsuperscript{48}. Different studies have shown that ML/DL algorithms are capable of predicting much more decently when classifying binary categories, compared to multiple categories. For example, a study by Senan et al.\textsuperscript{49} showed that a specific DL network achieved an accuracy of 99% and AUC of 97.5% for binary classification (COVID-19 vs healthy) compared to an accuracy and AUC of 95% and 97.1%, respectively, for classifying CXRs into COVID-19, viral pneumonia, lung opacity, and healthy individuals.

Regarding multi-class classification studies on COVID-19, some studies showed promising results\textsuperscript{50–52}. For instance, Wu et al.\textsuperscript{53} and Qian et al.\textsuperscript{54} evaluated the power of CXR-based and CT-based CNN models for differentiating between multiple classes of patients, including COVID-19, viral pneumonia, bacterial pneumonia, and healthy individuals, respectively. In addition to CNN models, some studies investigated multi-class categorization
power of ML models. For example, Hussain et al. assessed COVID-19, bacterial, viral, and healthy CXRs using extracted features and five ML algorithms. These algorithms classified each CXR into one of the four aforementioned CXR categories. They reached an accuracy and AUC of 0.79 and 0.87, respectively. A study by Khan et al. evaluated CT-based ML algorithms, such as multi-class SVM. In a recent study by Moradi Khanabi et al., two-step ML algorithms were proposed for diagnosis and severity scoring from COVID-19 CT images. They performed three-class classification for two different diagnostic tasks (normal, other pneumonia, and COVID-19 pneumonia) and severity scoring (mild, moderate and severe). They extracted radiomic features form whole lungs and used multiple machine learning algorithms for feature selection and classification purposes. They reported 0.909 ± 0.026, 0.907 ± 0.056, and 0.982 ± 0.010 for precision, recall, and AUC for diagnostic purposes and 0.868 ± 0.123 precision, 0.865 ± 0.121 recall, and 0.969 ± 0.022 AUC for severity scoring using a random forest algorithm.

Homayounieh et al. included 315 patients in their study and extracted CT-based radiomics features from the lungs to show that radiomics can predict patients’ outcome (inpatient vs. outpatient management) with an AUC of 0.84 while the radiologist assessment alone achieved an AUC of 0.69. Feature extraction was performed by applying the different preprocessing algorithms on images, with classification performed using logistic regression. They reported that adding clinical variables to the radiomics model can notably improve the predictability of a model for patient outcome prediction (AUC improved from 0.75 to 0.84). Another study conducted by Wei et al. evaluated the predictive ability of two models (one CT texture-based and one clinical) for determining the severity of each of the 81 COVID-19 patients. They showed that CT texture features could modestly predict whether the patient has common COVID-19 pneumonia or a severe one with an AUC of 0.93, which is comparable to that of the clinical-only model (AUC = 0.95). They also observed that several texture features had a moderate correlation with the clinical variables of patients.

Chaganti et al. studied Ground Glass Opacity (GGO) and consolidations that appear on a CT image of COVID-19 patients in an attempt to propose an automated method for segmenting and quantifying COVID-19 lesions. Their proposed method calculated the percentage of opacity and lung severity score using deep learning algorithms and was able to predict the severity with a decent performance. However, Chaganti et al. proposed a method trained only on the mentioned abnormalities and had a limited performance in other abnormalities quantification. Even with improving segmentation algorithms, this method would be limited because of the highly heterogeneous nature of COVID-19 pneumonia in addition to ignoring the shape and texture of segmented lesions. Moreover, providing accurate lobe segmentation of COVID-19 patients would be challenging from typical low-dose and high pitch chest CT scans. In the current and previous studies, radiomics features, as extracted from the entire lung (less challenging segmentation task for deep learning algorithms), were evaluated to provide fast and robust severity scoring in COVID-19 patients.

In this work, chest CT was used for assessment. At the same time, there are few studies on other modalities such as chest X-ray radiography in prognostication and outcome prediction evaluation of COVID-19 patients. For example, Bae and colleagues utilized radiomics features and modeled them on chest X-rays of 514 patients and found out that their radiomics- and deep learning-based model can accurately predict mortality and the need for mechanical ventilation in patients (AUCs = 0.93 and 0.90, respectively). Providing a severity score using chest X-rays is a valuable venue to explore. Yet, such work requires extensive comparisons with CT-based frameworks to assess the relative value of each modality for different tasks.

A number of radiomic features were selected with different IVs by two different algorithms. The 10Percentile, 90 Percentile and Mean from first-order features, which show the different percentile and Mean intensity within a region of interest were the selected features. The 10Percentile, 90 Percentile and Mean from first-order, despite max and min intensity, which are affected by noise, could be correlated with the involvement of the lung by infection as in severe cases, the infected lungs have high HU values. Other features selected by both algorithms was the Mean Absolute Deviation from first-order, achieving the highest IV in both algorithms. This feature is defined as the mean distance of image intensities from the mean value. As different stages of Covid-19 disease had different CT manifestations from no lesions, and medium to highly affected by infectious lesions, this feature could be correlated by stage of disease with different levels of infection demonstrated by the intensity of HUs. In addition to our study, these three features were selected by Yip et al. using the same datasets with different machine learning algorithms.

Zone Entropy (ZE) from GLSZM was another radiomics feature selected with high IV. This feature measures the randomness in distribution of the zones where a higher value indicates higher heterogeneity. Different stages of COVID-19 indicate different manifestations, including bilateral, multifocal, peripheral ground-glass opacities, consolidation, and crazy paving. These manifestations provide different textures where ZE could potentially be correlated with initial different heterogeneity generated by different stages. Dependence Variance (DV) from GLDM which measures the intensities variance had the highest IV in BRF algorithms. This feature could potentially be correlated with heterogeneities in different scores as severe cases had multiple types of lesions with high heterogeneity across the whole lung.

Gray Level Non-Uniformity Normalize from GLRLM was selected by both algorithms with high IV, which represents the spatial intensity changes in images. In severe COVID-19 cases, the lungs reveal more infections containing different types of manifestations resulting in high heterogeneity textures. In the case of high variability of intensity and high spatial change, such as high severe cases, the GLNUN feature value would be high. Gray Level Variance (GLV) from GLDM was another feature selected by MARS algorithm as high IV. GLDM calculates the coarseness of the texture whereas GLV feature measures the variance in dependence counts over intensities. This feature also quantifies the heterogeneity of regions of interest. In our study, this could be correlated with severe cases as the lung involves coarse textures of infection manifestation.

We presented the voxelwise feature map for three different features in different classes of severity. These features map visualize the different patterns of features across the different COVID-19 cases. In this study, we
attempted to clinically interpret selected features, similar to previous studies\textsuperscript{61-64} with the aim to hypothetically correlate the selected features and biological phenomena in different classes of severity. We should note that multivariate analysis uses different information from the selected features and using only one feature as univariate analysis doesn’t yield high performance for scoring. The combination of these selected features could provide complementary information toward robust multiclass severity scoring modeling.

This study suffered from a few limitations, including the fact that our model was trained on single-center data. Further research should be conducted on large-scale and multi-centric data and patient images with multiple observers for improved training of the models and enhanced generalizability. In the current study, the developed models were compared only to previous studies. Further work should focus on the comparison of ML-based scoring models with conventional scoring approaches.

Conclusion

We evaluated high-dimensional multinomial multiclass severity scoring of pneumonia using CT radiomic features and machine learning algorithms. We applied two feature selectors coupled to a classifier on a large cohort of COVID-19 patients. Our radiomics model was validated to depict accurate classification of patients according to multi-class pneumonia severity assessment criteria, highlighting the potential of this emerging paradigm in the assessment and management of COVID-19 patients. The selected radiomic features could be visualized to highlight the affected regions for better understanding of images, toward interpretable machine learning models. We proposed radiomics and machine learning-based high-dimensional multinomial multiclass severity scoring systems which could be potentially used in real clinical situations for severity assessment of COVID-19 patients. The proposed methods could be useful for highly affected (severe) COVID-19 patients management (ICU admission and treatment assessment).

Received: 31 May 2022; Accepted: 23 August 2022
Published online: 01 September 2022

References

1. Meyerowitz Katz, G. & Merone, L. A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. Int. J. Infect. Dis. 101, 138–148. https://doi.org/10.1016/j. ijid.2020.09.1464 (2020).
2. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C. & Di Napoli, R. In StatPears (StatPears Publishing Copyright © 2020, StatPears Publishing LLC., 2020).
3. Corman, V. M. et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. https://doi.org/10.2807/1560-7917.es.2020.25.7.2000045 (2020).
4. La Marca, A. et al. Testing for SARS-CoV-2 (COVID-19): A systematic review and clinical guide to molecular and serological in-vitro diagnostic assays. Reprod. Biomed. Online 41, 483–499. https://doi.org/10.1016/j.rbmo.2020.06.001 (2020).
5. Schmidt, C. W. CT scans: Balancing health risks and medical benefits. Environ. Health Perspect. 120, A118–A121. https://doi.org/10.1289/ehp.120-a118 (2012).
6. Fjerstad, M., Trussell, J., Lichtenberg, E. S., Sivin, I. & Cullins, V. Severity of infection following the introduction of new infection control measures for medical abortion. Contraception 83, 330–335. https://doi.org/10.1016/j.contraception.2010.08.022 (2011).
7. Tai, S. Y. & Xia, L. Coronavirus disease 2019 (COVID-19): Role of chest CT in diagnosis and management. AJR Am. J. Roentgenol. 214, 1280–1286. https://doi.org/10.2214/ajr.20.22954 (2020).
8. Li, Y. & Xia, L. Coronavirus disease 2019 (COVID-19): rRT-PCR or CT? Eur. J. Radiol. 126, 108961. https://doi.org/10.1016/j.ejrad.2020.108961 (2020).
9. Long, C. et al. Diagnosis of the coronavirus disease (COVID-19): rRT-PCR or CT? Eur. J. Radiol. 126, 108961. https://doi.org/10.1016/j.ejrad.2020.108961 (2020).
10. Tai, S. Y. et al. Symptom severity of patients with advanced cancer in palliative care unit: Longitudinal assessments of symptoms improvement. BMJ Palliat. Care 15, 32. https://doi.org/10.1186/s12904-016-0105-8 (2016).
11. Piarat, C., Pozzi, M., Croft, R., Carlisle, M. & Trappe, S. Severity scoring systems in the critically ill. Crit. Care 18, 181–185. https://doi.org/10.1093/cc/ccaapjx (2014).
12. Alhajeri, A. & Erwin, S. Acute pancreatitis: Value and impact of CT severity index. Abdom. Imaging 44, 1218–1223. https://doi.org/10.1007/s00261-020-01336-6 (2019).
13. Li, X. et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J. Allergy Clin. Immunol. 146, 110–118. https://doi.org/10.1016/j.jaci.2020.04.006 (2020).
14. Sanders, D. B., Li, Z., Brody, A. S. & Farrell, P. M. Chest computed tomography scores of severity are associated with future lung disease progression in children with cystic fibrosis. Am. J. Respir. Crit. Care Med. 184, 816–821. https://doi.org/10.1164/rccm.201105-0816OC (2011).
15. Sahb, B. et al. Severity assessment of acute pancreatitis using CT severity index and modified CT severity index: Correlation with clinical outcomes and severity grading as per the Revised Atlanta Classification. Indian J. Radiol. Imaging 27, 152–160. https://doi.org/10.4103/irji.IRJI_160_17 (2017).
16. Raghuvanshi, S., Gupta, R., Vyas, M. M. & Sharma, R. CT evaluation of acute pancreatitis and its prognostic correlation with CT severity index. J. Clin. Diagn. Res. 10, TC06–TC11. https://doi.org/10.7860/jcdr/2016/10949.7934 (2016).
17. Alhajeri, A. & Erwin, S. Acute pancreatitis: Value and impact of CT severity index. Abdom. Imaging 33, 18–20. https://doi.org/10.1007/s00261-007-9315-0 (2008).
18. Prokop, M. et al. CO-RADS: A categorical CT assessment scheme for patients suspected of having COVID-19-definition and evaluation. Radiology 296, E97-e104. https://doi.org/10.1148/radiol.202020147310.1148/radiol.202021473 (2020).
19. Neri, E. et al. Structured reporting of chest CT in COVID-19 pneumonia: A consensus proposal. Insights Imaging 11, 92. https://doi.org/10.1186/s13244-020-00901-7 (2020).
20. Francone, M. et al. Chest CT score in COVID-19 patients: Correlation with disease severity and short-term prognosis. Eur. Radiol. 30, 6808–6817. https://doi.org/10.1007/s00330-020-07033-y (2020).
21. Zhao, W., Zhong, Z., Xie, X., Yu, Q. & Liu, J. Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: A multicenter study. AJR Am. J. Roentgenol. 214, 1072–1077. https://doi.org/10.2214/ajr.20.22976 (2020).
22. Li, K. et al. The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. Invest. Radiol. 55, 327–331. https://doi.org/10.1097/rlr.0000000000001672 (2020).
23. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images are more than pictures, They are data. Radiology 278, 563–577. https://doi.org/10.1148/radiol.2015151169 (2016).
24. Arian, F. et al. Myocardial function prediction after coronary artery bypass grafting using MRI radiomics features and machine learning algorithms. J. Digit. Imaging in press (2022).
24. Khodabakhshi, Z., Shiri, I., Andratschke, N. & Tanadini-Lang, S. Two-year overall survival prediction in non–small-cell lung cancer patients using pre-treatment computed tomography images and deep neural networks: A multicentric study. Med. Imaging Deep. Learn. in press (2022).

25. Manafi-Farid, R. et al. [(18)F]FDG-PET/CT radiomics and artificial intelligence in lung cancer: Technical aspects and potential clinical applications. Semin. Nucl. Med. https://doi.org/10.1053/j.semnuclmed.2022.04.004 (2022).

26. Liu, Z. et al. The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges. Thera- nostics 9, 1303–1322. https://doi.org/10.7150/thno.30309 (2019).

27. Yip, S. S. F. & Aerts, H. J. W. L. Applications and limitations of radiomics. Phys. Med. Biol. 61, R150–R166. https://doi.org/10.1088/0031-9155/61/13/R150 (2016).

28. Cunliffe, A. et al. (18)F]FDG-PET/CT radiomics and artificial intelligence in lung cancer: Technical aspects and potential clinical applications. Semin. Nucl. Med. https://doi.org/10.1053/j.semnuclmed.2022.04.004 (2022).

29. Shiri, I. et al. Two-step machine learning to diagnose and predict involvement of lungs in COVID-19 and pneumonia using CT radiomics. medRxiv. https://doi.org/10.1101/2020.09.07.20189977 (2020).

30. Morozov, S. P. et al. MedsmEd: Chest CT scans with COVID-19-related findings dataset. arXiv preprint arXiv:2005.06465 (2020).

31. Jin, C. et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat. Commun. 11, 5088. https://doi.org/10.1038/s41467-020-18685-1 (2020).

32. Di, D. et al. COVID-19 prognostic modeling using CT radiomic features and machine learning algorithms: Analysis of a multi-institutional dataset of 14,339 patients. Comput. Biol. Med. 145, 105467. https://doi.org/10.1016/j.compbiomed.2022.105467 (2022).

33. Shiri, I. et al. Two-step machine learning to diagnose and predict involvement of lungs in COVID-19 and pneumonia using CT radiomics. medRxiv. https://doi.org/10.1101/2022.06.15.22276090 (2022).

34. Shiri, I. et al. High-dimensional multinomial multiclass severity scoring of COVID-19 pneumonia using CT radiomics features and machine learning algorithms. medRxiv. https://doi.org/10.1101/2020.04.27.22274369 (2022).

35. Bouchareb, Y. et al. Artificial intelligence-driven assessment of radiological images for COVID-19. Comput. Biol. Med. 104665 (2021).

36. Homayounieh, F. et al. Computed tomography radiomics can predict disease severity and outcome in coronavirus disease 2019 pneumonia. J. Comput. Assist. Tomogr. 44, 640–646. https://doi.org/10.1097/act.0000000000001094 (2020).

37. Wang, T. et al. Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: A longitudinal study. Radiology 296, E534–E544. https://doi.org/10.1148/rg.2020200843 (2020).

38. Yip, S. S. F. et al. Performance and robustness of machine learning-based radiomic COVID-19 severity prediction. medRxiv. https://doi.org/10.1101/2020.09.07.20189977 (2020).

39. Singh, A. K., Kumar, A., Mahmud, M., Kaiser, M. S. & Kishore, A. COVID-19 infection detection from chest X-ray images using hybrid social group optimization and support vector classifier. Cognit. Comput. https://doi.org/10.1007/s12559-021-09848-3 (2021).

40. Qad, T. S. et al. Hybrid deep-learning and machine-learning models for predicting COVID-19. Comput. Intell. Neurosci. https://doi.org/10.1155/2021/12083893 (2021).

41. Zwanenburg, A. et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295, 328–338 (2020).

42. van Griethuysen, J. J. M. et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 77, e104–e107 (2017).

43. Zhang, W., Wu, C., Li, Y., Wang, L. & Samui, P. Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. Georisk Assess. Manag. Risk Eng. Geohazards 15, 1–14 (2019).

44. Salimi, Y. et al. Deep learning-based fully automated Z-axis coverage range definition from scout scans to eliminate overscanning in chest CT imaging. Insights Imaging 12, 162. https://doi.org/10.1186/s13244-021-01105-3 (2021).

45. Khodabakhshi, Z. et al. Non-small cell lung carcinoma histopathological subtype typing by high-dimensional immunohistochemical multiclass CT radiomics signature. Comput. Biol. Med. 136, 104752 (2021).

46. Senan, E., Raghavan, M. Y., Alisharif, N. & Al-alhany, T. H. H. Automated diagnosis of chest X-ray for early detection of COVID-19 disease. Comput. Math. Methods Med. 6919483–6919483, 2021. https://doi.org/10.1155/6919483 (2021).

47. Singh, A. K., Kumar, A., Mahmud, M., Kaiser, M. S. & Kishore, A. COVID-19 infection detection from chest X-ray images using hybrid social group optimization and support vector classifier. Cognit. Comput. https://doi.org/10.1007/s12559-021-09848-3. https://doi.org/10.1038/s41598-022-18994-z
61. Amini, M. et al. Overall survival prognostic modelling of non-small cell lung cancer patients using positron emission tomography/computed tomography harmonised radiomics features: The quest for the optimal machine learning algorithm. Clin. Oncol. (R. Coll. Radiol.) 34, 114–127. https://doi.org/10.1016/j.clon.2021.11.014 (2022).

62. Shiri, I. et al. Impact of feature harmonization on radiogenomics analysis: Prediction of EGFR and KRAS mutations from non-small cell lung cancer PET/CT images. Comput. Biol. Med. 142, 105230. https://doi.org/10.1016/j.compbiomed.2022.105230 (2022).

63. Khodabakhshi, Z. et al. Overall survival prediction in renal cell carcinoma patients using computed tomography radiomic and clinical information. J. Digit. Imaging 34, 1086–1098. https://doi.org/10.1007/s10278-021-00500-y (2021).

64. Amini, M., et al. Multi-level multi-modality (PET and CT) fusion radiomics: Prognostic modeling for non-small cell lung carcinoma. Phys. Med. Biol. https://doi.org/10.1088/1361-6560/ac287d (2021).

Acknowledgements
This work was supported by the Swiss National Science Foundation under grant SNRF 320030_176052.

Author contributions
All authors contributed to drafting the manuscript and approved the final version.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to H.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022