Article

Machine learning based multi-parameter modeling for prediction of post-inflammatory lung changes

1

|  |
| --- |
| **Citation:** To be added by editorial staff during production.  Academic Editor: Firstname Lastname  Received: date  Revised: date  Accepted: date  Published: date    **Copyright:** © 2024 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). |

1 Department of Radiology, Medical University Innsbruck, Anichstrasse 35, 6020 Innsbruck, Austria.

**Abstract:**

Purpose:   
Prediction of lung function deficits following pulmonary infection with computed tomography (CT) findings, demographic, clinical and biochemical explanatory factors, and characteristic of acute infection is challenging and suffers from inaccuracy. IArtificial intelligence has transformed medical data analysis and may improve multi-parameter modeling. We sought to develop machine learning models for prediction of post-inflammatory lung changes during COVID-19 recovery.

Material and Methods:   
In the prospective CovILD study (n = 420 longitudinal observations from n = 140 COVID-19 survivors), data on lung function testing, chest CT including severity scoring by a human radiologist and density measurement by a software, demography, and persistent symptoms were collected. Those data sets were used to develop and evaluate machine learning models of of lung function test abnormalities and numeric readouts during post-inflammatory convalescence.

Results/Conclusion:   
Multi-parameter machine learning trained with demographic, clinical, and chest CT data reliably and reproducibly predicts functional lung deficits, and outperforms single markers of lung pathology and human radiologist’s assessment. It may improve diagnostic and foster personalized treatment.

**Keywords:** Covid, Artificial Intelligence, Lung CT, Quantification

1. Introduction

Pulmonary computed tomography (CT) has emerged as a cornerstone in the assessment of diverse lung pathologies, ranging from early-stage malignancies to chronic obstructive pulmonary disease (COPD), interstitial lung diseases (ILDs), and infection, such as COVID-19 pneumonia. Traditional approaches tof evaluation of pulmonary CT scans have predominantly relied on manual assessment by radiologists, a process characterized by subjectivity, time intensity, and inter-observer variability (1). The recent integration of software-based CT quantification techniques has offered a multifaceted arsenal of advantages over conventional scoring techniques, allowing for rapid and consistent analysis of voluminous imaging datasets (2). Recent studies have demonstrated the potential of AI algorithms to accurately detect and quantify COVID-19-related inflammatory lung changes on CT scans (3, 4). As such, CT quantification may not only facilitate risk stratification but may also contribute to our understanding of disease progression and treatment response (5). Furthermore, it may facilitate the identification of novel imaging biomarkers indicative of disease severity, prognosis, and therapeutic efficacy (6).

However, the prediction of lung function deficits following COVID-19 suffers from low accuracy (12). Additionally, relevance of residual structural lung lesions in CT and severity of lung damage in COVD-19 convalescents for pulmonary function is not entirely clear (9,11).

Machine learning based multi-parameter modeling may overcome some limitations of univariate correlations and receiver operating curve (ROC) analysis. The purpose of this paper was to develop and evaluate machine learning multi-parameter models with demographic, clinical, biochemical and CT quantitative data for prediction of post-inflammatory lung changes in patients who had COVID-19.

2. Materials and Methods  
  
Study data

Data of clinical and cardiopulmonary recovery recorded in the prospective multicenter CovILD study described in more detail in (9-12) (Medical University of Innsbruck, Austria (approval number: 1103/2020); ClinicalTrials.gov (NCT04416100)). A cohort of n = 145 COVID-19 survivors recruited between March and June 2020 at three clinical centers in Tyrol, Austria, was investigated two, three, six, and twelve months after diagnosis. . The study inclusion criteria were age ≥ 18 years, SARS-CoV-2 positivity confirmed by PCR and presence of COVID-19 symptoms. All participants were infected with the wild-type form of SARS-CoV-2. Herein, n = 420 longitudinal observations from 140 participants were analyzed, with complete CT of the chest and lung function testing (LFT) as analysis inclusion criteria. The severity of lung lesions in CT was scored by thoracic radiologists with a CT severity score (CTSS). In addition, an AI-based software (Syngo.via CT Pneumonia Analysis Software, Siemens Healthineers, Erlangen, Germany) was used to quantify lung density as percentage of the lung with opacity and high opacity (consolidation) (9, 10).

–**Analysis** outcomes and endpoints  
The following numeric LFT parameters were analyzed as percentage of patient’s reference values: DLCO (diffusion capacity for carbon monoxide), FVC (forced vital capacity), and FEV1 (forced expiratory volume in the first one second). Functional lung abnormalities were defined as DLCO < 80%, FVC <80%, and FEV1 < 80% of the reference.  
The primary analysis endpoint was construction and evaluation of multi-parameter models of numeric LFT parameters and lung function abnormalities with baseline and longitudinal demographic, clinical, and CT variables as explanatory factors.The primary endopoint was addressed by a machine learning modeling approach.  
  
The secondary analysis endpoint was assessment of human-determined CTSS and AI-measured CT lung opacity and high opacity as standalone predictors of LFT abnormalities and numeric LFT readouts. This endpoint was addressed by statistical hypothesis testing, correlation and receiver-operating characteristic (ROC) analysis.

Statistical analysis  
Details of statistical analysis are provided in Supplementary Methods.  
Statistical analysis was performed with R version 4.2.3 (R Foundation for Statistical Computing). Differences of CT readouts of lung damage severity were compared between donor-matched observations with and without reduced DLCO by a participant-wise blocked bootstrap test, and effect size was measured by biseral r effect size statistic (15). Cutoffs of CTSS, opacity and high opacity for detection of LFT abnormalities were found by maximizing the Youden’s J statistic. ROC analysis statistics (area under the curve [AUC], sensitivity, specificity, Cohen’s κ) for those optimal cutoffs were computed and their 95% confidence intervals were obtained by blocked bootstrap (16,17).  
DLCO, FVC and FEV1 (regression models), as well as abnormalities of DLCO, FVC (classification models) and FEV1 were modeled with 37 explanatory variables. The explanatory variables included time after diagnosis, demographic features (e.g. age, sex, body mass index, smoking, comorbidity), characteristic of acute COVID-19 (severity, medication) and recovery (e.g. weight loss, symptoms of relevance for lung function, time after diagnosis), and presence of human- and software-rated structural lung abnormalities in CT scans (e.g. GGO, CTSS, opacity and high opacity) ( Supplementary Table S2). The models were constructed with four common machine learning algorithms: Random Forest (18,19), gradient boosted machines (GBM) (10–22), neural network with a single hidden layer (23), and support vector machines (SVM) with radial kernel (24,25). Selection of the optimal values of parameters controlling model behavior (tuning) motivated by the maximum of Youden’s J statistic (classification models) or minimum mean absolute error (MAE, regression models) in 10-repeats 10-fold cross-validation (26). Because of the presence of participant-matched observation, patient-wise blocked cross-validation design was used both in the model selection and model evaluation. Model predictions were evaluated both in the training data and blocked 10-repeats 10-fold cross-validation. Predictions of the classification models were evaluated by accuracy, Cohen’s κ, ROC metrics (AUC, sensitivity, specificity), and Brier scores (16,2728). Predictions of the regression models were assessed pseudo-R2, MAE, andSpearman’s ρ coefficient of correlation between the predicted and observed values. Importance of explanatory variables was estimated by absolute values of SHAP statistics (Shapley additive explanations) (29–31).

3. Results

Machine learning modeling of lung function **durging COVID-19 recovery**   
Among the investigated LFT abnormalities, solely reduced DLCO defined as values < 80% of the patient’s reference was predicted with satisfactory accuracy in the cross-validation setting independently of the machine learning algorithm (overall accuracy: 0.82 to 0.85, κ: 0.45 to 0.5, AUC: 0.87 to 0.9). Low Brier score values for the models of reduced DLCO indicated good overall calibration (Brier score: 0.11 to 0.14). In turn, models of reduced FVC and FEV1 exhibited poor predictive performance in the cross-validation setting (overall accuracy: 0.72 to 0.81, κ: 0.094 to 0.17, AUC: 0.57 to 0.69, Figure 1A, Table 1, Supplementary Figure S9).

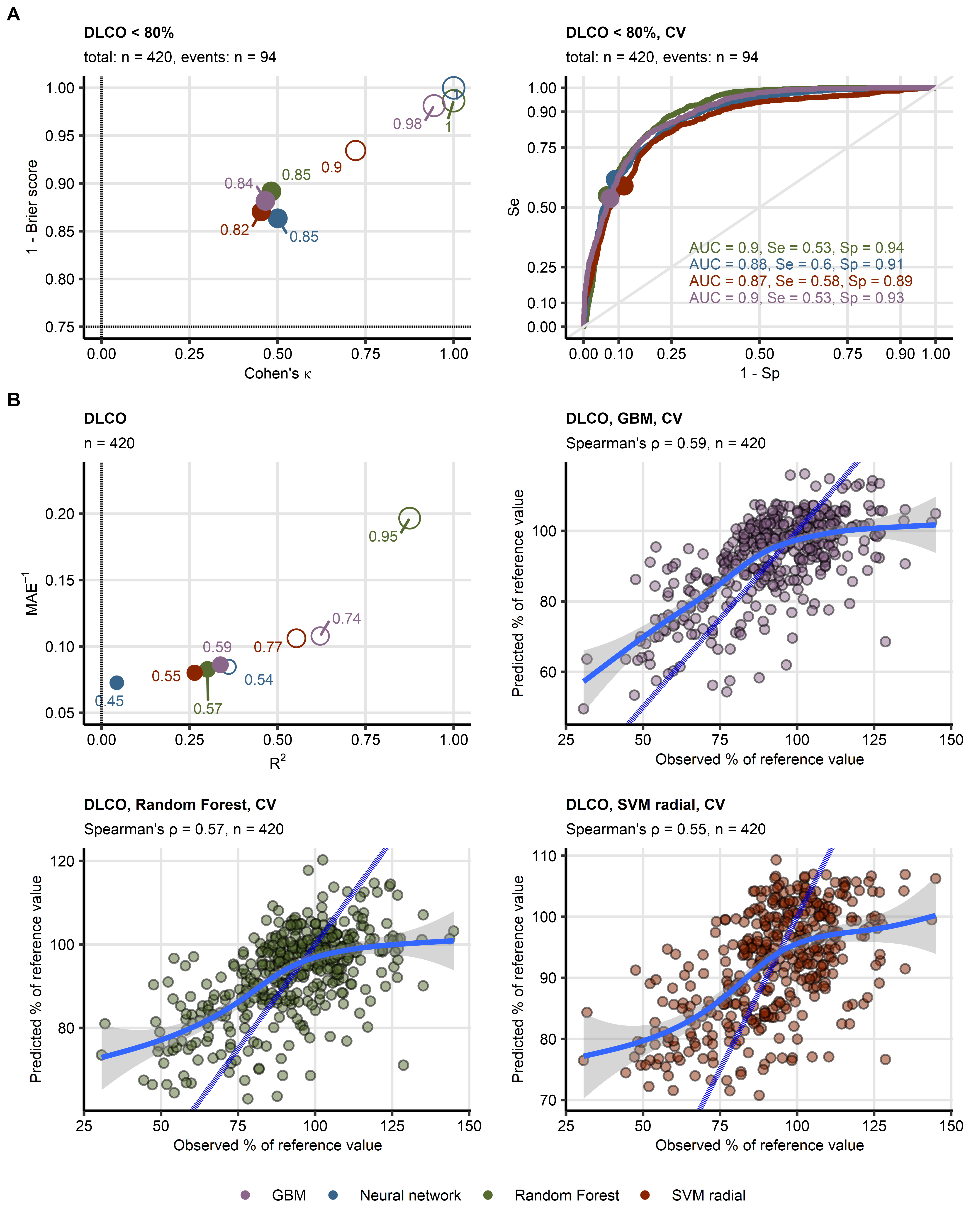
As in case of the LFT abnormalities, solely DLCO expressed as percentages of the patient’s reference was modeled with low error and substantial explanatory power in cross-validation with the Random Forest, GBM, and SVM algorithms (MAE: 11.6 to 12.5, R2: 0.26 to 0.34). Those models were also characterized by good calibration assessed by correlation between the predicted and observed DLCO (ρ: 0.55 to 0.59). By contrast, the neural network model of DLCO suffered from large error and lpoor explanatory peformance in the cross-validation setting (MAE = 13.8, R2 = 0.043). No meaningful models of FVC or FEV1 could be developed with any of the machine learning algorithms (cross-validation, R2: -0.086 to -0.03, Figure 1B, Table 2, Supplementary Figure S10).  
Abnormalities in lung CT quantified by the software (opacity, high opacity) and human radiologists (damage severity measured by CT severity score [CTSS], ground glass opacities [GGO], broncheictasis, reticulation) belonged to the 15 most influential explanatory variables for prediction of reduced DLCO by all four machine learning algorithms identified by Shapley additive explanations (Figure 2).

**Human-determined CTSS or software-determined density quantification as standalone markers of lung function impairment**

Human-determined CTSS (difference of medians [Δ]: 9 points, p < 0.001, effect size: r = 0.57), software-determined opacity (Δ: 1.3% of the lung, p < 0.001, effect size: r = 0.63) and high opacity (Δ: 0.063% of the lung, p < 0.001, effect size: r = 0.58) were significantly higher in observations with reduced DLCO than in the remaining data points with large effect sizes  
In a ROC analysis, software determined opacity with a cutoff of 0.12% of the lung exhibited substantially better AUC and sensitivity at detection of reduced DLCO (AUC: 0.81, sensitivity: 0.81, specificity: 0.69) than software-determined high opacity (cutoff: 0.002% of the lung, AUC: 0.79, sensitivity: 0.8, specificity: 0.68) and human-quantified lung damage (CTSS; cutoff: 4 points, AUC: 0.78, sensitivity: 0.78, specificity: 0.68, Figure 3, Table 3).

3.2. Figures, Tables and Schemes

Figure 1. Evaluation of performance of machine learning models of DLCO during COVID-19 convalescence.

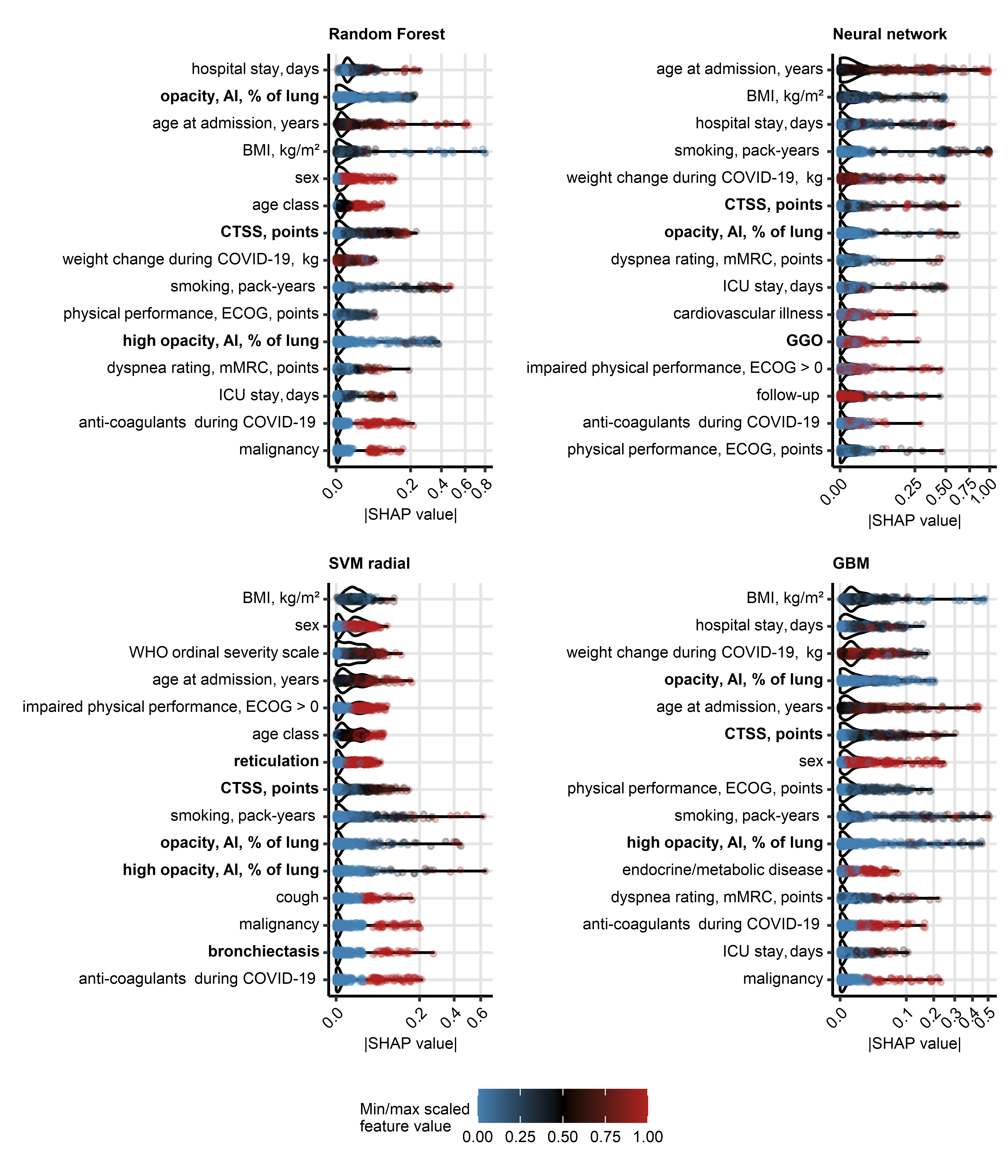


(A) Four machine learning classification models of insufficient diffusion capacity for carbon monoxide (DLCO < 80% of reference value: n = 94, total observations: n = 420) employing time after COVID-19 diagnosis, CT readouts, demographic and clinical explanatory variables were trained. Their performance was evaluated in the entire data set and 10-repeats 10-fold cross-validation with overall accuracy metric, Cohen’s κ as a measure of concordance between predicted and observed outcome, and Brier score as a calibration measure. Left: numeric performance measures of the models (open circles: the entire data set, filled circles: cross-validation); point sizes and point labels represent overall model accuracy, the dashed lines visualize values of κ and Brier score expected for prediction of insufficient DLCO be chance. Right: receiver-operating characteristic curves for predictions in cross-validation folds, numeric statistics are displayed in the plot. Numbers of complete observations and observations with DLCO insufficiency (‘events’) are displayed in the plot captions.

(B) Four machine learning regression models of DLCO (percentage of reference values, total observations: n = 420) employing time after COVID-19 diagnosis, CT readouts, demographic and clinical explanatory variables were trained. Their performance was evaluated in the entire data set and 10-repeats 10-fold cross-validation with R2 as a measure of explained variation, mean absolute error, and Spearman’s ρ coefficient of correlation between the predicted and observed values. Bubble plot: numeric performance measures of the model (open circles: the entire data set, filled circles: cross-validation); point sizes and point labels represent values of ρ, the dashed line visualizes the R2 value expected for a meaningless model. Scatter plots: observed and predicted values of diffusion capacity for carbon monoxide in cross-validation folds; the blue dashed lines with slope 1 and intercept 0 represent absolutely accurate predictions, general additive model trends with standard errors are visualized as the solid blue lines with gray ribbons. Numbers of complete observations and ρ valuesare displayed in the plot captions.

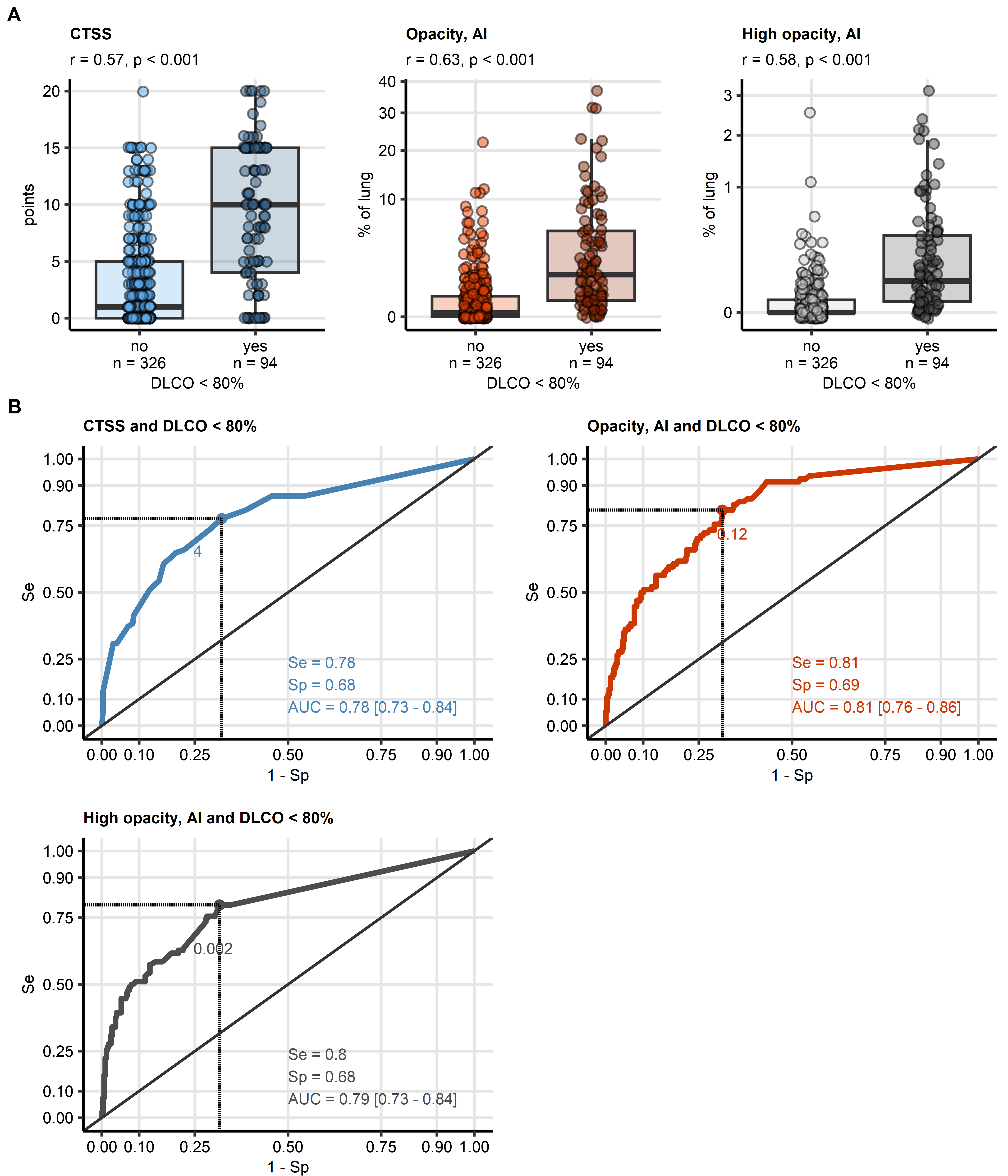
DLCO: diffusion capacity for carbon monoxide, CV: cross-validation; AUC: are under the receiver-operating characteristic curve; Se: sensitivity; Sp: specificity; MAE: mean absolute error; GBM: gradient boosted machines; SVM radial: support vector machines with radial kernel.

Figure 2. Explanatory variable importance for models of reduced DLCO measured by Shapley additive explanations.



Importance of explanatory variables for the machine learning models of reduced DLCO (< 80% of reference) was investigated by Shapley additive explanations (SHAP). Absolute SHAP values for explanatory variables with the 15 most influential variables are presented in violin plots. Points represent single observations, point colors code for minimum/maximum scaled value of the explanatory variable. Explanatory variables obtained via computed tomography are highlighted with bold font in the Y axes.  
CT: computed tomography; DLCO: diffusion capacity for carbon monoxide; opacity and high opacity, AI: opacity and high opacity of the lung determined by artificial intelligence; BMI: body mass index; CTSS: human-determined CT severity score, sum for all lobes; ECOG: Eastern Cooperative Oncology Group physical performance score; mMRC: modified Medical Research Council dyspnea scale; ICU: intensive care unit.

Figure 3. Detection of DLCO insufficiency by human- and artificial intelligence-determined CT readouts of severity of structural lung damage.



(A) Values of the radiological readouts of lung damage severity were compared between data points with and without insufficient diffusion capacity for carbon monoxide by blocked bootstrap test with r effect size statistic. Median values with interquartile ranges are depicted as boxes, whiskers span over 150% of the interquartile ranges, single observations are visualized as points. Effect sizes and p values are displayed in the plot captions. Numbers of observations are indicated in the X axis.

(B) Quality of detection of insufficient diffusion capacity for carbon monoxide with the radiological readouts of lung damage severity was assessed by receiver-operating characteristic (ROC) analysis. ROC curves are shown, the optimal cutoffs of the severity readuts determined by Youden’s criterion are represented by points with numbers. Sensitivity, specificity at the optimal cutoff, and area under the curve statistic with 95% confidence interval are displayed in the plots.

CT: computed tomography; DLCO: diffusion capacity for carbon monoxide; CTSS: human-determined CT severity score, sum for all lobes; opacity and high opacity, AI: opacity and high opacity of the lung determined by artificial intelligence; AUC: are under the curve of receiver-operating characteristic; Se: sensitivity; Sp: specificity.

**Tables:**

Table 1: Cross-validated performance of binary machine learning classifiers at predicting lung function testing (LFT) abnormalities.

| **Responsea** | **Algorithmb** | **Overall accuracyc** | **κd** | **Brier score** | **AUCe** | **Sensitivity** | **Specificity** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DLCO < 80% | Random Forest | 0.85 | 0.480 | 0.11 | 0.90 | 0.53 | 0.94 |
| Neural network | 0.85 | 0.500 | 0.14 | 0.88 | 0.60 | 0.91 |
| SVM radial | 0.82 | 0.450 | 0.13 | 0.87 | 0.58 | 0.89 |
| GBM | 0.84 | 0.470 | 0.12 | 0.90 | 0.53 | 0.93 |
| FVC < 80% | Random Forest | 0.79 | 0.110 | 0.16 | 0.69 | 0.14 | 0.95 |
| Neural network | 0.72 | 0.094 | 0.25 | 0.58 | 0.27 | 0.83 |
| SVM radial | 0.78 | 0.120 | 0.16 | 0.68 | 0.19 | 0.92 |
| GBM | 0.78 | 0.150 | 0.17 | 0.67 | 0.21 | 0.92 |
| FEV1 < 80% | Random Forest | 0.80 | 0.120 | 0.15 | 0.64 | 0.15 | 0.95 |
| Neural network | 0.75 | 0.110 | 0.21 | 0.57 | 0.26 | 0.86 |
| SVM radial | 0.80 | 0.130 | 0.16 | 0.59 | 0.18 | 0.94 |
| GBM | 0.81 | 0.170 | 0.16 | 0.61 | 0.21 | 0.94 |
| aLFT: lung function testing, DLCO: diffusion capacity for CO, FVC: forced vital capacity; FEV1: forced expiratory volume in one second. | | | | | | | |
| bSVM: support vector machines with radial kernel; GBM: gradient boosted machines. | | | | | | | |
| cRatio of correct predictions to the total observation number. | | | | | | | |
| dCohen κ statistic of inter-rater reliability between the predicted and observed outcome. | | | | | | | |
| eAUC: are under the receiver-operating characteristic curve. | | | | | | | |

Table 2: Cross-validated performance of regression machine learning models at predicting values of lung function testing parameters.

| **Responsea** | **Algorithmb** | **pseudo-R²c** | **MAEd** | **ρe** |
| --- | --- | --- | --- | --- |
| DLCO | Random Forest | 0.300 | 12 | 0.570 |
| Neural network | 0.043 | 14 | 0.450 |
| SVM radial | 0.260 | 12 | 0.550 |
| GBM | 0.340 | 12 | 0.590 |
| FVC | Random Forest | -0.030 | 10 | 0.220 |
| Neural network | -0.079 | 11 | 0.074 |
| SVM radial | -0.031 | 10 | 0.210 |
| GBM | -0.040 | 10 | 0.200 |
| FEV1 | Random Forest | -0.045 | 12 | 0.160 |
| Neural network | -0.086 | 12 | 0.210 |
| SVM radial | -0.039 | 11 | 0.190 |
| GBM | -0.047 | 12 | 0.170 |
| aDLCO: diffusion capacity for CO, FVC: forced vital capacity; FEV1: forced expiratory volume in one second. | | | | |
| bSVM: support vector machines with radial kernel; GBM: gradient boosted machines. | | | | |
| cDefined as 1 - ratio of mean squared error and variance. | | | | |
| dMAE: mean absolute error. | | | | |
| eΡ: Spearman coefficient of correlation between the predicted and observed response values. | | | | |

Table 3: Detection of reduced diffusion capacity for CO DLCO < 80% reference value) by single CT-derived parameters: AI-determined opacity and high opacity, and human-determined CT severity score.

| **CT variablea** | **Cutoffb** | **Statisticc** | **Value, 95% CI** |
| --- | --- | --- | --- |
| CTSS |  | AUC | 0.78 [0.727 - 0.84] |
| 4.000 | κ | 0.34 [0.23 - 0.45] |
| 4.000 | Sensitivity | 0.78 [0.64 - 0.89] |
| 4.000 | Specificity | 0.68 [0.6 - 0.75] |
| high opacity, AI |  | AUC | 0.79 [0.734 - 0.84] |
| 0.002 | κ | 0.37 [0.26 - 0.47] |
| 0.002 | Sensitivity | 0.8 [0.7 - 0.89] |
| 0.002 | Specificity | 0.68 [0.62 - 0.75] |
| opacity, AI |  | AUC | 0.81 [0.763 - 0.86] |
| 0.120 | κ | 0.38 [0.27 - 0.48] |
| 0.120 | Sensitivity | 0.81 [0.72 - 0.89] |
| 0.120 | Specificity | 0.69 [0.62 - 0.75] |
| aCTSS: human-determined CT severity score, sum for al lung lobes; high opacity and opacity, AI: percentage of the lungs with high opacity and opacity determined by artificial intelligence. | | | |
| bCutoff of the CT variable corresponding to the maximum of Jouden Y statistic. | | | |
| cAUC: area under the curve of receiver-operating characteristic; κ: Cohen κ statistic of inter-rater reliability between the predicted and observed outcome, computed for the CT variable cutoff; sensitivity and specificity: sensitivity and specificity computed at the CT variable cutoff. | | | |

4. Discussion

With four unrelated machine learning algorithms, Radom Forest, SVM, GBM, and neural networks, we developed accurate models of reduced DLCO < 80% in patients after COVID-19. The highest concordance between the predicted and observed DLCO < 80% was achieved for moderate COVID-19 convalescents at the two- to six-month follow-up examinations (all algorithms, κ: 0.45, 0.69). The accuracy was in turn the poorest for ambulatory patients at the six- and twelve-month follow-up (all algorithms, κ: 0, 0.46, Supplementary Figure S11A). The Random Forest, SVM and GBM models of DLCO expressed as percentage of the patient’s reference exhibited the lowest average errors for moderate COVID-19 survivors (mean error: -2.8 to 2.9). In turn, DLCO values were systematically overestimated for severe COVID-19 patients at the two- to six-month follow-up visits (mean error: 1.2 to 6.5) and underestimated for ambulatory COVID-19 convalescents at the same time points (mean error: -4.2 to 0.12). The best performance in the hospitalized moderate subset is likely attributed to the large number of observations and frequent DLCO deficits in this group of participants available for training of the models (n = 47 observations of DLCO < 80% in 234 observations from moderate COVID-19 patients). Conversely, the numbers of observations from ambulatory COVID-19 patients and numbers of observations DLCO insufficiency in this group was low (10 observations of DLCO < 80% in 85 observations from ambulatory patients) which resulted in relatively poor performance of the machine learning models.

As investigated by absolute values of SHAP variable importance metrics (29,30), human- and AI-derived ratings of structural lung damage belonged to the most influential explanatory variables of the models of DLCO < 80% and the meaningful models of DLCO. Other influential explanatory variables were well characterized risk factors of severe COVID-19 (age, male sex, body mass index, pre-existing malignancy and cardiovascular disease), readouts of acute COVID-19 severity (severity class, WHO ordinal severity scale, hospitalization length, ICU treatment, anti-coagulant treatment, and weight change), smoking intensity (pack-years), as well as rating of physical performance impairment (ECOG) and exertional dyspnea (mMRC).

Because CTSS, lung opacity, and high opacity were identified to be crucial for prediction of reduced DLCO <80% and DLCO by machine learning, we explored the association of the human- and AI-determined readouts of lung damage with DLCO in more detail. AI-measured opacity of the lung, but not CTSS or high opacity, was found to be significantly higher in observations with reduced FVC (difference of medians: 0.4, p = 0.0051, effect size: r = 0.29) and insufficient FEV1 (difference of medians: 0.3, p = 0.009, effect size: r = 0.27); sizes of those effects were, however, small (Supplementary Table S10). Furthermore, CTSS, opacity and high opacity correlated negatively with moderate effect size with DLCO (p < 0.001, Spearman’s ρ: -0.46 to -0.44). Interestingly, the CT readouts of structural lung damage also correlated significantly with FVC (Spearman’s ρ: -0.3 to -0.21) and FEV1 (Spearman’s ρ: -0.27 to -0.14); effect size of those associations was small (Supplementary Table S11).

Finally, in an univariable ROC analysis, human-determined CTSS (AUC = 0.78, 95% CI: 0.73 to 0.84), and AI-determined lung opacity (AUC = 0.81, 95% CI: 0.76 to 0.86) and high opacity (AUC = 0.79, 95% CI: 0.73 to 0.84) were identified as standalone markers of insufficient DLCO in COVID-19 patients. The optimal cutoffs of CTSS, opacity, and high opacity for detection of DLCO < 80% were, respectively, 4 points, 0.12% of the lung, and 0.002% of the lung, and allowed for identification of insufficient DLCO with moderate sensitivity and specificity (sensitivity: 0.78 to 0.81, specificity: 0.68 to 0.69, Figure 4B). In case of the AI measures of lung opacity and high opacity, the extremely low values of the optimal cutoffs let us infer that event low-grade radiological abnormalities can be associated with clinically relevant functional lung impairment. Yet, concordance between the predicted observed DLCO insufficiency assessed by Cohen’s κ for CTSS, opacity, and high opacity as standalone markers (blocked bootstrap κ: 0.34 to 0.38, Table 4) was substantially lower that in the multi-parameter machine learning models (cross-validated κ: 0.45 to 0.5). This underlines the importance of other CT-unrelated explanatory factors such acute disease course for reliable prediction of functional lung deficits in COVID-19 convalescents.

Limitations

Our study has several limitations. First, the overall patient and observation number was low, in particular for ambulatory COVID-19 convalescents. Analogically, the study cohort was enriched in hospitalized individuals with in constitute a minute fraction of COVID-19 patients in the real world setting. Second, complete sets of longitudinal CT and LFT measurements at the two-, three-, six- and twelve-month follow-up examinations were available solely for 55 patients. Especially the number of observations obtained with ambulatory and moderate COVID-19 convalescents at the six- and twelve-month follow-up were substantially lower as compared with earlier time points. The incompleteness of the longitudinal data may have hence compromised performance of the machine learning models in particular for ambulatory COVID-19 cases and at the later time points. Third, because of the limited number of participants and observations, we abstained from definition of a test subset of the data used solely for bias-free model evaluation (hold-out strategy). Instead, both model selection and evaluation was done with blocked repeated cross-validation, which may have overestimated performance of the models. Hence, external validation of our findings in an independent cohort is recommended. Finally, the analyzed cohort was recruited in the initial phase of the pandemic and consisted of individuals infected with the wild-type variant of the SARS-CoV-2 virus. For this reason, it is not completely clear, how our findings translate to the recent variants of the pathogen and how the pulmonary recovery is affected by anti-SARS-CoV-2 immunity, improved treatment and care. However, it is feasible, that the CT severity readouts, human-determined CTSS as well as AI-determined opacity and high opacity, are equally applicable in the post-pandemic setting as standalone markers of functional lung impairment during recovery from COVID-19 and other respiratory infections.

5. Conclusions

AI-based multi-parameter models outperformed univariable correlations and ROC analyses for prediction in predicting functional lung deficits. CT imaging data including CTSS, lung opacity, and high opacity were identified to be crucial for prediction of reduced DLCO <80%. AI-based multi-parameter modeling has the potential to improve outcome prediction and guide personalized treatment.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.” Please turn to the [CRediT taxonomy](https://img.mdpi.org/data/contributor-role-instruction.pdf) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** This research received no external funding

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section “MDPI Research Data Policies” at https://www.mdpi.com/ethics.

**Conflicts of Interest:** The authors declare no conflicts of interest.

References

1. Huang L, Han R, Ai T, Yu P, Kang H, Tao Q, Xia L. Serial Quantitative Chest CT Assessment of COVID-19: A Deep Learning Approach. Radiol Cardiothorac Imaging. 2020 Mar 30;2(2):e200075. doi: 10.1148/ryct.2020200075. PMID: 33778562; PMCID: PMC7233442.

2. Mei X, Lee HC, Diao KY, Huang M, Lin B, Liu C, Xie Z, Ma Y, Robson PM, Chung M, Bernheim A, Mani V, Calcagno C, Li K, Li S, Shan H, Lv J, Zhao T, Xia J, Long Q, Steinberger S, Jacobi A, Deyer T, Luksza M, Liu F, Little BP, Fayad ZA, Yang Y. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med. 2020 Aug;26(8):1224-1228. doi: 10.1038/s41591-020-0931-3. Epub 2020 May 19. PMID: 32427924; PMCID: PMC7446729.

3. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 2020 Jun;121:103792. doi: 10.1016/j.compbiomed.2020.103792. Epub 2020 Apr 28. PMID: 32568675; PMCID: PMC7187882.

4. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shen D, Shi Y. Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction. Med Phys. 2021 Apr;48(4):1633-1645. doi: 10.1002/mp.14609. Epub 2021 Mar 9. PMID: 33225476; PMCID: PMC7753662.

5. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Dahly DL, Damen JAA, Debray TPA, de Jong VMT, De Vos M, Dhiman P, Haller MC, Harhay MO, Henckaerts L, Heus P, Kammer M, Kreuzberger N, Lohmann A, Luijken K, Ma J, Martin GP, McLernon DJ, Andaur Navarro CL, Reitsma JB, Sergeant JC, Shi C, Skoetz N, Smits LJM, Snell KIE, Sperrin M, Spijker R, Steyerberg EW, Takada T, Tzoulaki I, van Kuijk SMJ, van Bussel B, van der Horst ICC, van Royen FS, Verbakel JY, Wallisch C, Wilkinson J, Wolff R, Hooft L, Moons KGM, van Smeden M. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020 Apr 7;369:m1328. doi: 10.1136/bmj.m1328. Update in: BMJ. 2021 Feb 3;372:n236. Erratum in: BMJ. 2020 Jun 3;369:m2204. PMID: 32265220; PMCID: PMC7222643.

6. Jacobs C, van Rikxoort EM, Twellmann T, Scholten ET, de Jong PA, Kuhnigk JM, Oudkerk M, de Koning HJ, Prokop M, Schaefer-Prokop C, van Ginneken B. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. Med Image Anal. 2014 Feb;18(2):374-84. doi: 10.1016/j.media.2013.12.001. Epub 2013 Dec 17. PMID: 24434166.

7. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44-56. doi: 10.1038/s41591-018-0300-7. Epub 2019 Jan 7. PMID: 30617339.

8. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, Tang A. Deep Learning: A Primer for Radiologists. Radiographics. 2017 Nov-Dec;37(7):2113-2131. doi: 10.1148/rg.2017170077. PMID: 29131760.

9. Sahanic S, Tymoszuk P, Luger AK, Hüfner K, Boehm A, Pizzini A, Schwabl C, Koppelstätter S, Kurz K, Asshoff M, et al. COVID-19 and its continuing burden after 12 months: a longitudinal observational prospective multicentre trial. ERJ open research (2023) 9:00317–2022. doi: 10.1183/23120541.00317-2022

10. Sonnweber T, Sahanic S, Pizzini A, Luger A, Schwabl C, Sonnweber B, Kurz K, Koppelstätter S, Haschka D, Petzer V, et al. Cardiopulmonary recovery after COVID-19: An observational prospective multicentre trial. European Respiratory Journal (2021) 57: doi: 10.1183/13993003.03481-2020

11. Luger AK, Sonnweber T, Gruber L, Schwabl C, Cima K, Tymoszuk P, Gerstner AK, Pizzini A, Sahanic S, Boehm A, et al. Chest CT of Lung Injury 1 Year after COVID-19 Pneumonia: The CovILD Study. Radiology (2022) 304:462–470. doi: 10.1148/radiol.211670

12. Sonnweber T, Tymoszuk P, Sahanic S, Boehm A, Pizzini A, Luger A, Schwabl C, Nairz M, Grubwieser P, Kurz K, et al. Investigating phenotypes of pulmonary COVID-19 recovery: A longitudinal observational prospective multicenter trial. eLife (2022) 11: doi: 10.7554/ELIFE.72500

13. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: Glossary of terms for thoracic imaging. (2008) 246:697–722. doi: 10.1148/radiol.2462070712

14. Ripley B. MASS: Support Functions and Datasets for Venables and Ripley’s MASS. (2022) https://cran.r-project.org/package=MASS

15. Mangiafico S. rcompanion: Functions to Support Extension Education Program Evaluation. (2022) https://cran.r-project.org/package=rcompanion

16. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica (2012) 22:276. doi: 10.11613/bm.2012.031

17. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F. Optimalcutpoints: An R package for selecting optimal cutpoints in diagnostic tests. Journal of Statistical Software (2014) 61:1–36. doi: 10.18637/jss.v061.i08

18. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software (2017) 77:1–17. doi: 10.18637/JSS.V077.I01

19. Breiman L. Random forests. Machine Learning (2001) 45:5–32. doi: 10.1023/A:1010933404324

20. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models. (2022) https://cran.r-project.org/package=gbm

21. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis (2002) 38:367–378. doi: 10.1016/S0167-9473(01)00065-2

22. Friedman JH. Greedy function approximation: A gradient boosting machine. https://doiorg/101214/aos/1013203451 (2001) 29:1189–1232. doi: 10.1214/AOS/1013203451

23. Ripley BD. Pattern recognition and neural networks. Cambridge University Press (2014). doi: 10.1017/CBO9780511812651

24. Weston J, Watkins C. Multi-Class Support Vector Machines. (1998)

25. Karatzoglou A, Hornik K, Smola A, Zeileis A. kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software (2004) 11:1–20. doi: 10.18637/JSS.V011.I09

26. Kuhn M. Building predictive models in R using the caret package. Journal of Statistical Software (2008) 28:1–26. doi: 10.18637/jss.v028.i05

27. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement (1960) 20:37–46. doi: 10.1177/001316446002000104

28. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Monthly Weather Review (1950) 78:1–3. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

29. Covert I, Lee SI. Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression. Proceedings of Machine Learning Research (2020) 130:3457–3465. https://arxiv.org/abs/2012.01536v3

30. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (2017) 2017-Decem:4766–4775. https://arxiv.org/abs/1705.07874v2

31. Mayer M, Watson D, Biecek P. kernelshap: Kernel SHAP. (2023) <https://cran.r-project.org/web/packages/kernelshap/index.html>

32. Gao M, Bagci U, Lu L, Wu A, Buty M, Shin HC, Roth H, Papadakis GZ, Depeursinge A, Summers RM, Xu Z, Mollura DJ. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Comput Methods Biomech Biomed Eng Imaging Vis. 2018;6(1):1-6. doi: 10.1080/21681163.2015.1124249. Epub 2016 Jun 6. PMID: 29623248; PMCID: PMC5881940.

33. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, Bussink J, Gillies RJ, Mak RH, Aerts HJWL. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. PLoS Med. 2018 Nov 30;15(11):e1002711. doi: 10.1371/journal.pmed.1002711. PMID: 30500819; PMCID: PMC6269088.

34. Jacobs C, van Rikxoort EM, Twellmann T, Scholten ET, de Jong PA, Kuhnigk JM, Oudkerk M, de Koning HJ, Prokop M, Schaefer-Prokop C, van Ginneken B. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. Med Image Anal. 2014 Feb;18(2):374-84. doi: 10.1016/j.media.2013.12.001. Epub 2013 Dec 17. PMID: 24434166.

35. Soussan M, Chouahnia K, Maisonobe JA, Boubaya M, Eder V, Morère JF, Buvat I. Prognostic implications of volume-based measurements on FDG PET/CT in stage III non-small-cell lung cancer after induction chemotherapy. Eur J Nucl Med Mol Imaging. 2013 May;40(5):668-76. doi: 10.1007/s00259-012-2321-7. Epub 2013 Jan 11. PMID: 23306807.

36. Xu, M., Qi, S., Yue, Y. et al. Segmentation of lung parenchyma in CT images using CNN trained with the clustering algorithm generated dataset. BioMed Eng OnLine 18, 2 (2019). https://doi.org/10.1186/s12938-018-0619-9