AI-Driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia

Guillaume Chassagnon, MD PhD[*,1,2,3], Maria Vakalopoulou, PhD[*4,5,6,7], Enzo Battistella, MsC[4,6,7], Stergios Christodoulidis, PhD[8,9], Trieu-Nghi Hoang-Thi, MD[1], Severine Dangeard, MD[1], Eric Deutsch, MD PhD[6,7], Fabrice Andre, MD PhD[8,9], Enora Guillo, MD[1], Nara Halm, MD[1], Stefany El Hajj, MD[1], Florian Bompard, MD[1], Sophie Neveu, MD[1], Chahinez Hani, MD[1], Ines Saab, MD[1], Aliénor Campredon, MD[1], Hasmik Koulakian, MD[1], Souhail Bennani, MD[1], Gael Freche, MD[1], Aurelien Lombard, MsC[15], Laure Fournier, MD PhD[2,10], Hippolyte Monnier, MD[10], Téodor Grand, MD[10], Jules Gregory, MD[2,11], Antoine Khalil, MD PhD[2,12], Elyas Mahdjoub, MD[2,12], Pierre-Yves Brillet, MD PhD[13], Stéphane Tran Ba, MD[13], Valérie Bousson, MD PhD[2,14], Marie-Pierre Revel, MD PhD[1,2,3], and Nikos Paragios, PhD[†4,7,15]

[1]Radiology Department, Hopital Cochin – AP-HP.Centre Université de Paris, 27 Rue du Faubourg Saint-Jacques, 75014 Paris, France
[2]Université de Paris, 85 boulevard Saint-Germain, 75006 Paris, France
[3]Inserm U1016, Institut Cochin, 22 rue Méchain, 75014 Paris, France
[4]Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systémes, Gif-sur-Yvette, France, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France
[5]Université Paris-Saclay, CentraleSupélec, Inria, Gif-sur-Yvette, France
[6]Université Paris-Saclay, Institut Gustave Roussy, Inserm 981 Molecular Radiotherapy and Innovative Therapeutics, 114 Rue Edouard Vaillant, 94800 Villejuif, France
[7]Gustave Roussy-CentraleSupélec-TheraPanacea, Noesia Center of Artificial Intelligence in Radiation Therapy and Oncology, Gustave Roussy Cancer Campus, Villejuif, France
[8]Université Paris-Saclay, Institut Gustave Roussy, Inserm 1030 Predictive Biomarkers and New Therapeutic Strategies in Oncology, 114 Rue Edouard Vaillant, 94800 Villejuif, France
[9]Université Paris-Saclay, Institut Gustave Roussy, Prism Precision Medicine Center, 114 Rue Edouard Vaillant, 94800 Villejuif, France
[10]Radiology Department, Hopital Européen Georges Pompidou – AP-HP.Centre Université de Paris, 20 Rue Université Paris-Saclay, 75015 Paris, France
[11]Radiology Department, Hopital Beaujon – AP-HP.Nord Université de Paris, 100 Boulevard du Général Leclerc, 92110 Clichy
[12]Radiology Department, Hopital Bichat – AP-HP.Nord Université de Paris, 46 Rue Henri Huchard, 75018 Paris, France
[13]Radiology Department, Hopital Avicenne – AP-HP.Hopitaux universitaires Paris Seine-Saint-Denis, 125 Rue de Stalingrad, 93000 Bobigny, France
[14]Radiology Department, Hopital Lariboisiére – AP-HP.Nord Université de Paris, 2 Rue Ambroise Paré, 75010 Paris, France
[15]TheraPanacea, 27 Rue du Faubourg Saint-Jacques, 75014 Paris, France

[*]Dr. Guillaume Chassagnon & Dr. Maria Vakalopoulou have equally contributed to this work
[†]Corresponding author: n.paragios@therapanacea.eu

ABSTRACT

Chest computed tomography (CT) is widely used for the management of Coronavirus disease 2019 (COVID-19) 49 pneumonia because of its availability and rapidity[1–3] . The standard of reference for confirming COVID-19 relies on microbiological tests but these tests might not be available in an emergency setting and their results are not immediately available, contrary to CT. In addition to its role for early diagnosis, CT has a prognostic role by allowing visually evaluating the extent of COVID-19 lung abnormalities[4,5]. The objective of this study is to address prediction of short-term outcomes, especially need for mechanical ventilation. In this multi-centric study, we propose an end-to-end artificial intelligence solution for automatic quantification and prognosis assessment by combining automatic CT delineation of lung disease meeting expert's performance and data-driven identification of biomarkers for its prognosis. AI-driven combination of variables with CT-based biomarkers offers perspectives for optimal patient management given the shortage of intensive care beds and ventilators[6,7].

Main

COVID-19 has emerged in December 2019 in the city of Wuhan in China[8] and disseminated around the world, leading the World Health Organization to declare the COVID-19 outbreak a pandemic. The disease is caused by the SARS-Cov-2 virus and the leading cause of death is respiratory failure due to severe viral pneumonia[9]. Chest computed tomography (CT) has rapidly gained a major role for COVID-19 diagnosis. Indeed, despite being considered as the gold standard to make a definitive diagnosis, reverse transcription polymerase chain reaction (RT-PCR) suffers from false negatives, shortage of available supply test kits and long turnaround times[10–12].

Artificial intelligence has gained significant attention during the past decade and many applications have been proposed in medical imaging, including segmentation and characterization tasks such as lung cancer screening on CT[13–15]. A few studies have already reported deep learning to diagnose COVID-19 pneumonia on chest radiograph[16] or CT[17] . Other authors used

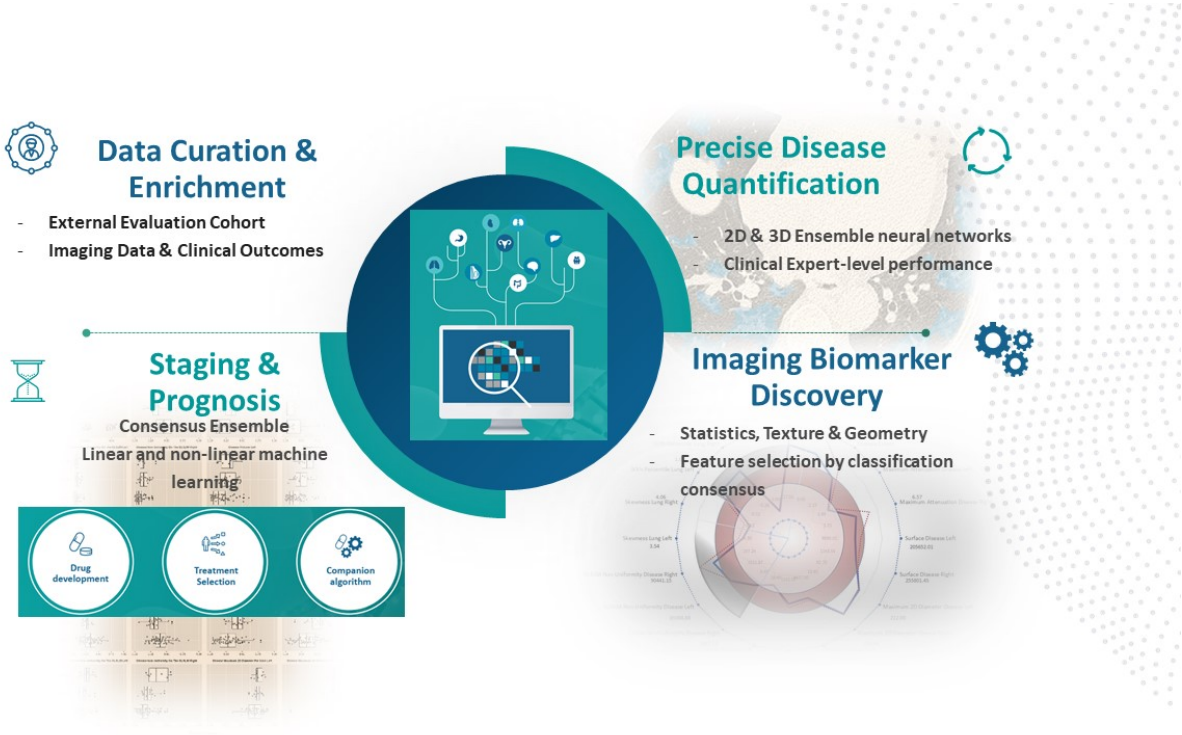
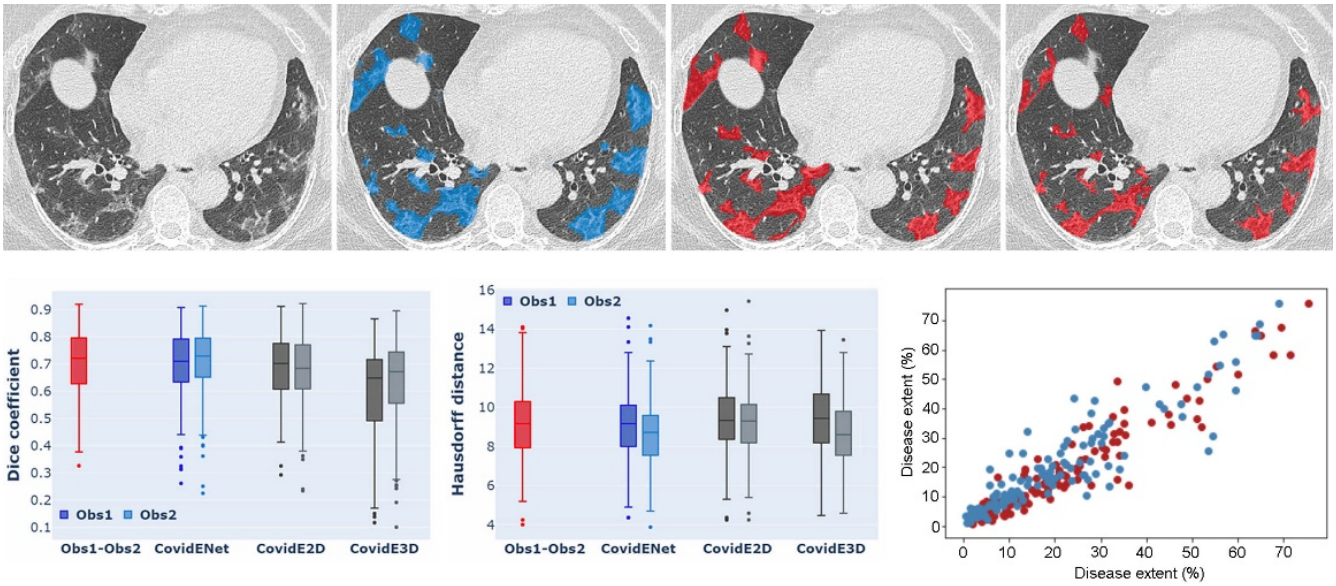


Figure 1. Overview of the method for CT-based quantification, staging and prognosis of COVID-19. (i) Two independent cohorts with quantification based on ensemble 2D & 3D consensus neural networks reaching expert-level annotations on massive evaluation, (ii) Consensus-driven bio(imaging)-marker selection on the principle of prevalence across methods leading to variables highly-correlated with outcomes & (iii) Consensus of linear & non-linear classification methods for staging and prognosis reaching optimal performance (minimum discrepancy between training & testing).



Method	Dice						Hausdorff Distance					
	Mean		Median		STD		Mean		Median		STD	
	Obs1	Obs2	Obs1	Obs2	Obs1	Obs2	Obs1	Obs2	Obs1	Obs2	Obs1	Obs2
CovidE2D	0.69	0.67	0.70	0.68	±0.13	±0.13	9.40	9.23	9.33	9.30	±1.83	±1.80
CovidE3D	0.62	0.65	0.67	0.70	±0.17	±0.16	9.43	8.70	9.43	8.60	±1.87	±1.81
CovidENet	0.69	0.70	0.71	0.73	±0.13	±0.13	9.18	8.75	9.16	8.72	±1.86	±1.78
Obs1 -Obs2	0.70		0.72		±0.12		9.16		9.16		±1.83	
CovidENet	**0.70**		**0.72**		**±0.12**		**8.96**		**8.94**		**±1.82**	

Figure 2. Comparison between automated and manual segmentations. Delineation of the diseased areas on chest CT in a COVID-19 patient: First Row: input, AI-segmentation, expert I-segmentation, expert II-segmentation. Second Row: Box-Plot Comparisons in terms of Dice similarity and Haussdorf between AI-solution, expert I & expert II, & Plot of correlation between disease extent automatically measured and the average disease extent measured from the 2 manual segmentations. Disease extent is expressed as the percentage of lung affected by the disease. Third row: statistical measures on comparisons between AI, expert I, and expert II segmentations.

deep-learning to quantify COVID-19 disease extent on CT but none of them used a multi-centric cohort while providing comparisons with segmentations done by radiologists[18,19]. Disease extent is the only parameter that can be visually estimated on chest CT to quantify disease severity[4,5], but visual quantification is difficult and usually coarse. Several AI-based tools have been recently developed to quantify interstitial lung diseases (ILD)[20–23], which share common CT features with COVID-19 pneumonia, especially a predominance of ground glass opacities. In this study, we investigated a fully automatic method (Figure 1) for disease quantification, staging and short-term prognosis. The approach relied on (i) a disease quantification solution that exploited 2D & 3D convolutional neural networks using an ensemble method, (ii) a biomarker discovery approach sought to determine the share space of features that are the most informative for staging & prognosis, & (iii) an ensemble robust supervised classification method to distinguish patients with severe vs non-severe short-term outcome and among severe patients those intubated and those who did not survive.

Part I: Disease Quantification

In the context of this work, we report a deep learning-based segmentation tool to quantify COVID-19 disease and lung volume. For this purpose, we used an ensemble network approach inspired by the AtlasNet framework[22]. We investigated a combination of 2D slice-based[24] and 3D patch-based ensemble architectures[25]. The development of the deep learning-based segmentation solution was done on the basis of a multi-centric cohort of 478 unenhanced chest CT scans (208,668 slices) of COVID-19

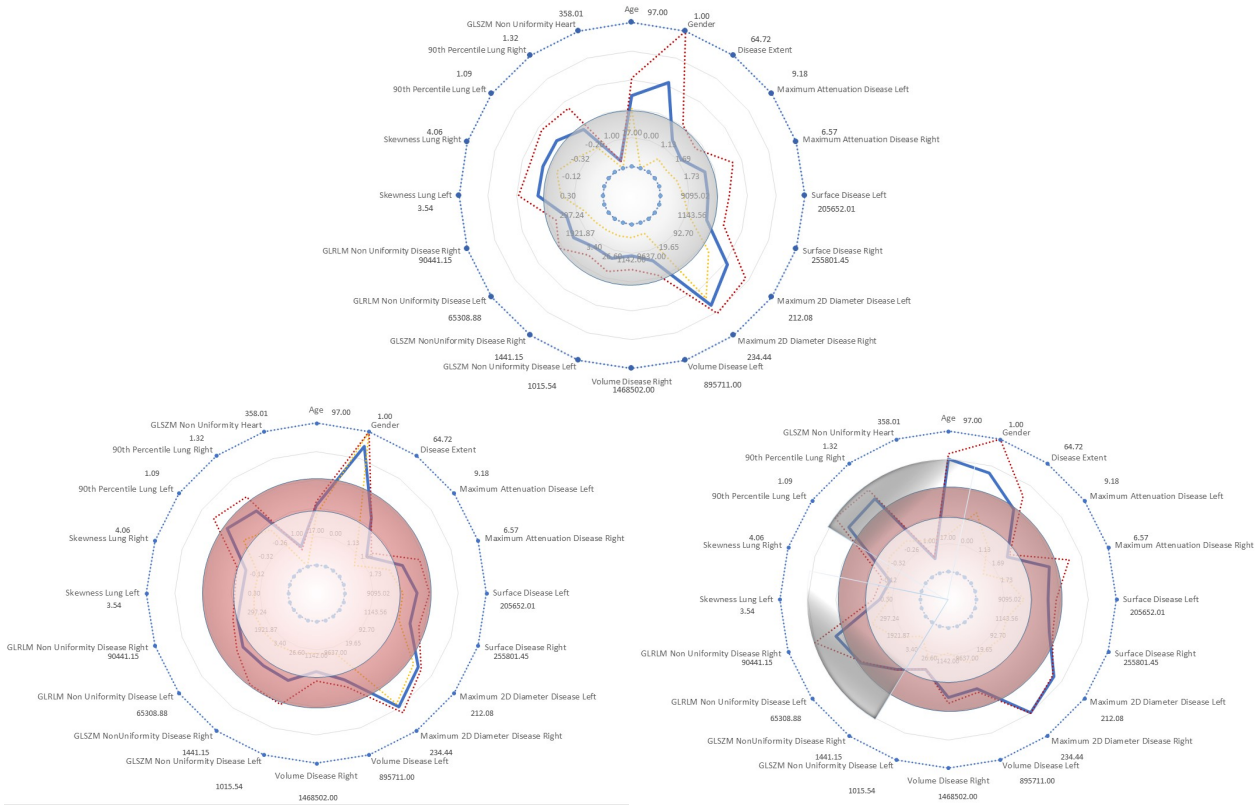
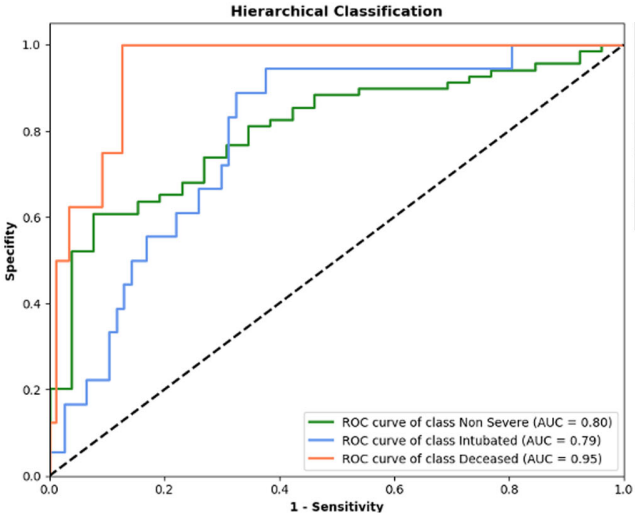


Figure 3. Spider-chart distribution of features depicting their minimum and maximum values [mean value (blue), 70% percentile (yellow) and 90% percentile (red) lines] with respect to the different outcomes with the following order: top: non-severe, bottom left: intensive care support & bottom right: deceased in the testing set. White and red circles represent respectively 40% and 60% of the maximum value of each feature. Clear separation was observed on these feature space with respect to the non-severe & severe cases. In terms of deceased versus intensive care patients, notable difference were observed with respect to three variables, the age of the patient, the condition of the healthy lung and the non-uniformity of the disease (indicated with gray in the spider-chart).

patients with positive RT-PCR. The multicentric dataset was acquired at 6 Hospitals, equipped with 4 different CT models from 3 different 91 manufacturers, with different acquisition protocols and radiation dose (Table 1). Fifty CT exams from 3 centers were used for training and 130 CT exams from 3 other centers were used for test (Table 2). Disease and lung were delineated on all $23,423$ images used as training dataset, and on only 20 images per exam but by 2 independent annotators in the test dataset ($2,600$ images). The overall annotation effort took approximately 800 hours and involved 15 radiologists with 1 to 7 years of experience in chest imaging. The consensus between manual (2 annotators) and automated segmentation was measured using the Dice similarity score (DSC)[26] and the Haussdorf distance (HD). The CovidENet performed equally well to trained radiologists in terms of DSCs and better in terms HD (Figure 2). The mean/median DSCs between the two expert's annotations on the test dataset were $0.70/0.72$ for disease segmentation. For the same task, DSCs between CovidENet and the manual segmentations were $0.69/0.71$ and $0.70/0.73$. In terms of HDs, the observed average value between the two experts was 9.16mm while it was 8.96mm between CovidENet and the two experts. When looking at disease extent, defined as the percentage of lung affected by the disease, we found no significant difference between automated segmentation and the average of the two manual segmentations ($19.9\% \pm 17.7$ $[0.5 - 73.2]$ vs $19.5\% \pm 16.5$ $[1.1 - 75.7]$; p= 0.352).

Part II: Imaging Biomarker Discovery

To assess the prognostic value of the Chest computed tomography (CT) an extended multi-centric data set was built. We reviewed outcomes in patient charts within the 4 days following chest CT and divided the patients in 3 groups: those who didn't survive, those who required mechanical ventilation and those who were still alive and not intubated. Out of the 478 included patients, 27 died (6%) and 83 were intubated (17%), forming a group of 110 patients with severe short-term outcome (23%). Data of 383 patients from 3 centers were used for training and those of 85 patients from 3 other centers composed



Classification Performance
(NS: Non-severe, **S:** severe (I&D), **I:** intubated, **D:** deceased**)**

Classifier	Balanced Accuracy		Weighted Precision		Weighted Sensitivity		Weighted Specificity	
	Training	Test	Training	Test	Training	Test	Training	Test
NS/I/D	0.79	**0.68**	0.82	**0.79**	0.68	**0.67**	0.84	**0.83**
NS/S	0.75	**0.74**	0.82	**0.79**	0.69	**0.69**	0.81	**0.79**
I/D	0.93	**0.81**	0.95	**0.90**	0.95	**0.88**	0.91	**0.74**

Confusion Matrices
(NS: Non-severe, **S:** severe (I&D), **I:** intubated, **D:** deceased**)**

NS vs S		I vs D				NS	I	D
65%	35%	100%	0		NS	65%	30%	5%
15%	85%	38%	62%		I	22%	78%	0%
					D	0%	38%	62%

Figure 4. Classification performance of dual & aggregated classifiers with respect to the non-severe vs severe case, the intubated vs deceased case and the three classes. Sensitivity and confusion tables are presented with respect to the different classification problems.

an independent test dataset (Table 3). Radiomics-based prognosis gained significant attention in the recent years towards predicting treatment outcomes[27]. In this study we have adopted a similar strategy, we extracted 107 features related to first order, higher order statistics, texture and shape information for lungs, disease extent and heart. Feature selection was performed on a basis of predictive value consensus. We created several representative partitions 117 of the training set (80% training and 20% validation) and run 13 different supervised classification methods towards optimal separation of the observed clinical ground truth between severe and non-severe cases (Table 4). The features that were shared between the different classifiers were retained as robust imaging biomarkers using a cut-off probability of 0.25 and were aggregated to patients' age and gender (Table 5). In total 12 features were retained for the prognosis part and included age, gender, disease extent, descriptors of disease heterogeneity and extension, features of healthy lung and a descriptor of cardiac heterogeneity. Correlations for some these features and the clinical outcome are presented in Figure 5 while a representation of these feature space with respect to the different classes is presented in Figure 3.

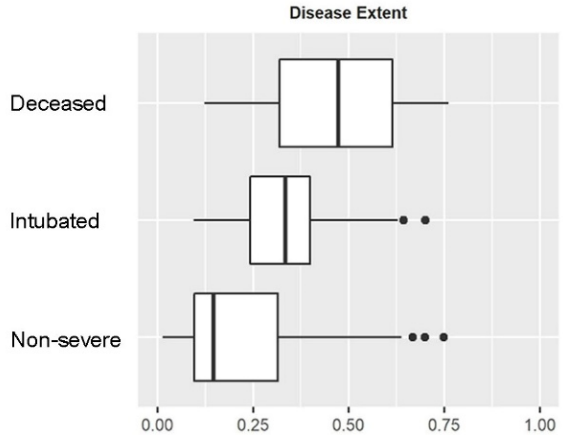
Part III: Staging and Prognosis

The staging/prognosis was implemented using a hierarchical classification principle, targeting first staging and subsequently prognosis. The staging component sought to separate patients with severe and non-severe short-term outcomes, while the prognosis sought to predict the risk of decease among severe patients. On the basis of the feature selection step, the machine learning algorithms that had a balanced accuracy greater than 60% on validation were considered. The selection of these methods was done on the basis of minimum discrepancy between performance on training and internal validation sub-training data set. We have built two sequential classifiers using this ensemble method, one to determine the severe cases and a second to predict survival. The classifier aiming to separate patients with severe and non-severe short-term outcomes had a balanced accuracy of 74%, a weighted precision of 79%, a weighted sensitivity of 69% and specificity of 79% to predict a severe short-term outcome (Figure 4, Table 6). The performance of the second classifier aiming to differentiate between intubated and deceased patients was even higher with a balanced accuracy of 81% (Figure 4, Table 7). The hierarchical classifiers combing the 3 classes had a balanced accuracy of 68%, a weighted precision of 79%, a weighted sensitivity of 67% and specificity of 83% (Figure 4). It was observed that prognosis performance difference between training and external cohort testing was low, suggesting that the most important information present at CT scans was recovered, and additional information should be integrated in order to fully explain the outcome.

Part IV: Conclusions

In conclusion, artificial intelligence enhanced the value of chest CT by providing fast accurate, and precise disease extent quantification and by helping to identify patients with severe short-term outcomes. This could be of great help in the current context of the pandemic with healthcare resources under extreme pressure. In a context where the sensitivity of RT-PCR has

Feature	Prev.	Cor.
G- Age	0.7	0.19
G- Gender	0.6	0.14
D- GLSZM Size Zone Non-Uniformity	0.8	0.30
D- First Order Maximum	0.8	0.18
D- Shape Voxel Volume	0.6	0.39
D- Extent	0.6	0.35
D- Shape Surface Area Disease	0.4	0.31
D- Shape Maximum 2D Diameter Slice	0.4	0.23
D- GLRLM Run Length Non-Uniformity	0.3	0.33
L- First Order Skewness	0.7	-0.34
L- First Order 90%	0.6	0.31
H- GLSZM Gray Level Non-Uniformity	0.7	0.15



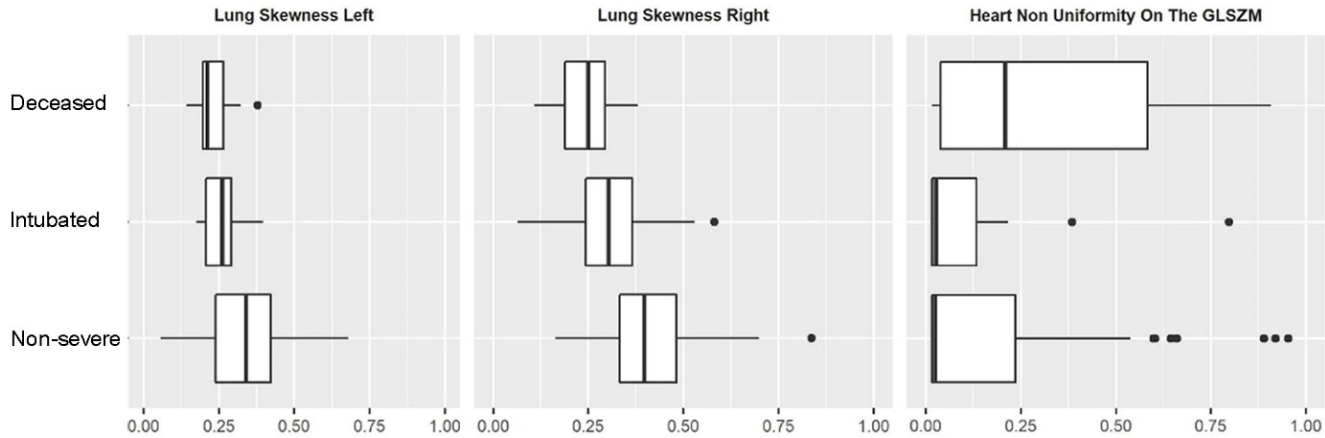


Figure 5. Discovery of – imaging-biomarkers through consensus. Generic variables (G: age, sex), disease related variables (D: extent, volume, maximum diameter, etc.), lung variables (L: skewness, etc.) as well as heart related variables (H: non-uniformity) have been automatically selected. The prevalence of the features as well as their distribution with respect to the different classes is presented for some of them, with rather clear separation and strong correlations with ground truth.

been shown to be low, such as 63% when perform on nasal swab[28], chest CT has been shown to provide higher sensitivity for diagnosis of COVID-19 as compared with initial RT-PCR from pharyngeal swab samples[10]. The current COVID-19 pandemic requires implementation of rapid clinical triage in healthcare facilities to categorize patients into different urgency categories[29], often occurring in the context of limited access to biological tests. Beyond the diagnostic value of CT for COVID-19, our study suggests that AI should be part of the triage process. The developed tool will be made publicly available. Our prognosis and staging method achieved state of the art results through the deployment of a highly robust ensemble classification strategy with automatic feature selection of imaging biomarkers and patients' characteristics available within the image' metadata. In terms of future work, the continuous enrichment of the data base with new examples is a necessary action on top of updating the outcome of patients included in the study. The integration of non-imaging data and other related clinical and categorical variables such as lymphopenia, the D-dimer level and other comorbidities[9,30–32] is a necessity towards better understanding the disease and predicting the outcomes. This is clearly demonstrated from the inability of any of the state-of-the art classification methods (including neural networks and multi-layer perceptron models) to predict the outcome with a balanced accuracy greater to 80% on the training data. Our findings could have a strong impact in terms of (i) patient stratification with respect to the different therapeutic strategies, (ii) accelerated drug development through rapid, reproducible and quantified assessment of treatment response through the different mid/end-points of the trial, and (iii) continuous monitoring of patient's response to treatment.

Methods

Study Design and Participants

This retrospective multi-center study was approved by our Institutional Review Board (AAA-2020-08007) which waived the need for patients' consent. Patients diagnosed with COVID-19 from March 4th to 29th at six large University Hospitals were eligible if they had positive PCR-RT and signs of COVID-19 pneumonia on unenhanced chest CT. A total of 478 patients formed the full dataset (208,668 CT slices). Only one CT examination was included for each patient. Exclusion criteria were (i) contrast medium injection and (ii) important motion artifacts.

For the COVID-19 radiological pattern segmentation part, 50 patients from 3 centers (A: 20 patients; B: 15 patients, C: 15 patients) were included to compose a training and validation dataset, 130 patients from the remaining 3 centers (D: 50 patients; E: 50 patients, F: 30 patients) were included to compose the test dataset (Table 2). The proportion between the CT manufacturers in the datasets was pre-determined in order to maximize the model generalizability while taking into account the data distribution.

For the radiomics driven prognosis study, 298 additional patients from centers A (96 patients), B (64 patients) and D (138 patients) were included to increase the size of the dataset. Data of 383 patients from 3 centers (A, B and D) were used for training and those of 85 patients from 3 other centers (C, E, F) composed an independent test set (Table 3). Only one CT examination was included for each patient. Exclusion criteria were (i) contrast medium injection and (ii) important motion artifacts. For short-term outcome assessment, patients were divided into 2 groups: those who died or were intubated in the 4 days following the CT scan composed the severe short-term outcome subgroup, while the others composed the non-severe short-term outcome subgroup.

CT Acquisitions

Chest CT exams were acquired on 4 different CT models from 3 manufacturers (Aquilion Prime from Canon Medical Systems, Otawara, Japan; Revolution HD from GE Healthcare, Milwaukee, WI; Somatom Edge and Somatom AS+ from Siemens Healthineer, Erlangen, Germany). The different acquisition and reconstruction parameters are summarized in Table 1. CT exams were mostly acquired at 120 (n=103/180; 57%) and 100 kVp (n=76/180; 42%). Images were reconstructed using iterative reconstruction with a 512×512 matrix and a slice thickness of 0.625 or 1 mm depending on the CT equipment. Only the lung images reconstructed with high frequency kernels were used for analysis. For each CT examination, dose length product (DLP) and volume Computed Tomography Dose Index (CTDIvol) were collected.

Data Annotation

Fifteen radiologists (GC, TNHT, SD, EG, NH, SEH, FB, SN, CH, IS, HK, SB, AC, GF and MB) with 1 to 7 years of experience in chest imaging participated in the data annotation which was conducted over a 2-week period. For the training and validation set for the COVID-19 radiological pattern segmentation, the whole CT examinations were manually annotated slice by slice using the open source software ITKsnap [1]. On each of the 23,423 axial slices composing this dataset, all the COVID-19 related CT abnormalities (ground glass opacities, band consolidations, and reticulations) were segmented as a single class. Additionally, the whole lung was segmented to create another class (lung). To facilitate the collection of the ground truth for the lung anatomy, a preliminary lung segmentation was performed with Myrian XP-Lung software (version 1.19.1, Intrasense, Montpellier, France) and then manually corrected.

As far as test cohort for the segmentation is concerned, 20 CT slices equally spaced from the superior border of aortic arch to the lowest diaphragmatic dome were selected to compose a 2,600 images dataset. Each of these images were systematically annotated by 2 out of the 15 participating radiologists who independently performed the annotation. Annotation consisted of manual delineation of the disease and manual segmentation of the lung without using any preliminary lung segmentation.

Deep Learning Construction

The segmentation tool was built under the paradigm of ensemble methods using a 2D fully convolutional network together with the AtlasNet framework[22] and a 3D fully convolutional network[25]. The AtlasNet framework combines a registration stage of the CT scans to a number of anatomical templates and consequently utilizes multiple deep learning-based classifiers trained for each template. At the end, the prediction of each model is - to the original anatomy and a majority voting scheme is used to produce the final projection, combining the results of the different networks. A major advantage of the AtlasNet framework is that it incorporates a natural data augmentation by registering each CT scan to several templates. Moreover, the framework is agnostic to the segmentation model that will be utilized. For the registration of the CT scans to the templates, an elastic registration framework based on Markov Random Fields was used, providing the optimal displacements for each template[33].

The architecture of the implemented segmentation models was based on already established fully convolutional neural network designs from the literature[24,25]. Fully convolutional networks following an encoder decoder architecture both in 2D

[1] http://www.itksnap.org

and 3D were developed and evaluated. For the 2D models the CT scans were separated on the axial view. The network included 5 convolutional blocks, each one containing two Conv-BN-ReLU layer successions. Maxpooling layers were also distributed at the end of each convolutional block for the encoding part. Transposed convolutions were used on the decoding part to restore the spatial resolution of the slices together with the same successions of layers. For the 3D pipeline, the model similarly consisted of five blocks with a down-sampling operation applied every two consequent Conv3D-BN-ReLU layers. Additionally, five decoding blocks were utilized for the decoding path, at each block a transpose convolution was performed in order to up-sample the input. Skip connections were also employed between the encoding and decoding paths. In order to train this model, cubic patches of size $64 \times 64 \times 64$ were randomly extracted within a close range of the ground truth annotation border in a random fashion. Corresponding cubic patches were also extracted from the ground truth annotation masks and the lung anatomy segmentation masks. To this end, we trained the model with the CT scan patch as input, the annotation patch as target and the lung anatomy annotation patch as a mask for calculating the loss function only within the lung region. In order to train all the models, each CT scan was normalized by cropping the Hounsfield units in the range $[-1024, 1000]$.

Regarding implementation details, 6 templates were used for the AtlasNet framework together with normalized cross correlation and mutual information as similarities metrics. The networks were trained using weighted cross entropy loss using weights depending on the appearance of each class and dice loss. Moreover, the 3D network was trained using a dice loss. The Dice loss (DL) and weighted cross entropy (WCE) are defined as follows,

$$DL = 1 - \frac{2pg+1}{p+g+1}, \qquad WCE = -(\beta g \log(p) + (1-g) log(1-p))$$

where p is the predicted from the network value and g the target/ ground truth value. β is the weight given for the less representative class. For network optimization, we used only the class for the diseased regions.

For the 2D experiments we used classic stochastic gradient descent for the optimization with initial learning rate = 0.01, decrease of learning rate = $2.5 \cdot 10^{-3}$ every 10 epochs, momentum =0.9 and weight decay =$5 \cdot 10^{-4}$. For the 3D experiments we used the AMSGrad and a learning rate of 0.001. The training of a single network for both 2D and 3D network was completed in approximately 12 hours using a GeForce GTX 1080 GPU, while the prediction for a single CT scan was done in a few seconds. Training and validation curves for one template of AtlasNet and the 3D network are shown in Figure 6. Both Dice similarity score and Haussdorff distances were higher with the 2D approach compared to the 3D approach (Figure **??**). However, the combination of their probability scores led to a significant improvement. Thus, the ensemble of 2D and 3D architectures was selected for the final COVID-19 segmentation tool.

Moreover, segmentation masks of the lung and heart of all patients were extracted by using ART-Plan software (Thera-Panacea, Paris, France). ART-Plan is a CE-marked solution for automatic annotation of organs, harnessing a combination of anatomically preserving and deep learning concepts. This software has been trained using a combination of a transformation and an image loss. The transformation loss penalizes the normalized error between the prediction of the network and the affine registration parameters depicting the registration between the source volume and the whole body scanned. These parameters are determined automatically using a downhill simplex optimization approach. The second loss function of the network involved an image similarity function – the zero-normalized cross correlation loss – that seeks to create an optimal visual correspondence between the observed CT values of the source volume and the corresponding ones at the full body CT reference volume. This network was trained using as input a combination of $360,000$ pairs of CT scans of all anatomies and full body CT scans. These projections used to determine the organs being present on the test volume. Using the transformation between the test volume and the full body CT, we were able to determine a surrounding patch for each organ being present in the volume. These patches were used to train the deep learning model for each full body CT. The next step consisted of creating multiple annotations on the different reference spaces, and for that a 3D fully convolutional architecture was trained for every reference anatomy. This architecture takes as input the annotations for each organ once mapped to the reference anatomy and then seeks to determine for each anatomy a network that can optimally segment the organ of interest similar to the AtlasNet framework used for the disease segmentation. This information was applied for every organ of interest presented in the input CT Scan. In average, $6,600$ samples were used for training per organ after data augmentation. These networks were trained using a conventional dice loss. The final organ segmentation was achieved through a winner takes all approach over an ensemble networks approach. For each organ, and for each full body reference CT a specific network was built, and the segmentation masks generated for each network were mapped back to the original space. The consensus of the recommendations of the different subnetworks was used to determine the optimal label at the voxel level.

Staging / Prognosis

As a preprocessing step, all images were resampled by cubic interpolation to obtain isometric voxels with sizes of 1 mm. Subsequently, disease, lung and heart masks were used to extract 107 radiomic features[34] for each of them (left and right lung

were considered separately both for the disease extent and entire lung). The features included first order statistics, shape-based features in 2D and 3D together with texture-based features. Radiomics features were enriched with clinical data available from the image metadata (age, gender), disease extent and number of diseased regions. The minimum and maximum values were calculated for the training and validation cohorts and Min-Max normalization was used to normalize the features, the same values were also applied on the test set. As a first step, a number of features were selected using a lasso linear model in order to decrease the dimensionality. The lasso estimator seeks to optimize the following objective function:

$$\frac{||y - Xw||_2^2}{2n} + \alpha ||w||_1$$

where α is a constant, $||w||_1$ is the L1-norm of the coefficient vector and n is the number of samples. The Lasso method was used with 200 alphas along a regularization path of length 0.01 and limited to 1000 iterations. The staging/prognosis component was addressed using an ensemble learning approach. First, the training data set was subdivided into training and validation set on the principle of $80\% - 20\%$ while respecting that the distribution of classes between the two subsets was identical to the observed one. We have created 10 subdivisions on this basis and evaluated the average performance of the following supervised classification methods: Nearest Neighbor, {Linear, Sigmoid, Radial Basis Function, Polynomial Kernel} Support Vector Machines, Gaussian Process, Decision Trees, Random Forests, AdaBoost, Gaussian Naive Bayes, Bernoulli Naive Bayes, Multi-Layer Perceptron & Quadratic Discriminant Analysis. Features selection was performed on the training set of each subdivision. In particular, the features selected in at least three subdivision were considered critical and have been used later for the staging and prognosis.

- Age
- Gender
- Disease Extent
- From the diseased areas: maximum attenuation, surface, maximum 2D diameter per slice, volume, non-uniformity of the Gray level Size Zone matrix (GLSZM) and non-uniformity of the Gray level Run Length matrix (GLRLM)
- From the lung areas: skewness and 90th percentile
- From the area of the heart: non-uniformity on the Gray level Size Zone matrix (GLSZM)

These features included first order features (maximum attenuation, skewness and 90th percentile), shape features (surface, maximum 2D diameter per slice and volume) and texture features (non-uniformity of the GLSZM and GLRLM).

Subsequently, this reduced feature space was considered to be most appropriate for training, and the following 7 classification methods with acceptable performance, $> 60\%$ in terms of balanced accuracy, as well as coherent performance between training and validation, performance decrease $< 20\%$ for the balanced accuracy between training and validation, were trained and combined together through a winner takes all approach to determine the optimal outcome (Table 4). The final selected methods include the {Linear, Polynomial Kernel, Radial Basis Function} Support Vector Machines, Decision Trees, Random Forests, AdaBoost, and Gaussian Naive Bayes which were trained and combined together through a winner takes all approach to determine the optimal outcome. To overcome the unbalance of the different classes, each class received a weight inversely proportional to its size. The Support Vector Machines were all three granted a polynomial kernel function of degree 3 and a penalty parameter of 0.25. In addition, the one with a Radial Basis Function kernel was granted a kernel coefficient of 3. The decision tree classifier was limited to a depth of 3 to avoid overfitting. The random forest classifier was composed of 8 of such trees. AdaBoost classifier was based on a decision tree of maximal depth of 2 boosted three times.

The classifiers were applied in a hierarchical way, performing first the staging and then the prognosis. More specifically, a majority voting method was applied to classify patients into severe and non-severe cases (Table 6). Then, another majority voting was applied on the cases predicted as severe only to classify them into intubated or deceased (Table 7). In such a setup, the correlation of the reported features are summarized in Table 5. For the hierarchical prognosis on the three classes a voting classifier for the prediction of each class against the others has been applied to aggregate the predicted outcomes from the 7 selected methods. In the Figure 8 we visualize the distributions of the different features along the ground truth labels and the prediction of the hierarchical classifier for each subject. In particular, all the samples are grouped using their ground truth labels and a boxplot is generated for each group and each feature. Additionally, color coded points are over imposed at each boxplot denoting the prediction label. It is therefore clearly visible that some features such as the disease extent, the age, the shape of the disease and the uniformity seems to be very important on separating the different subjects.

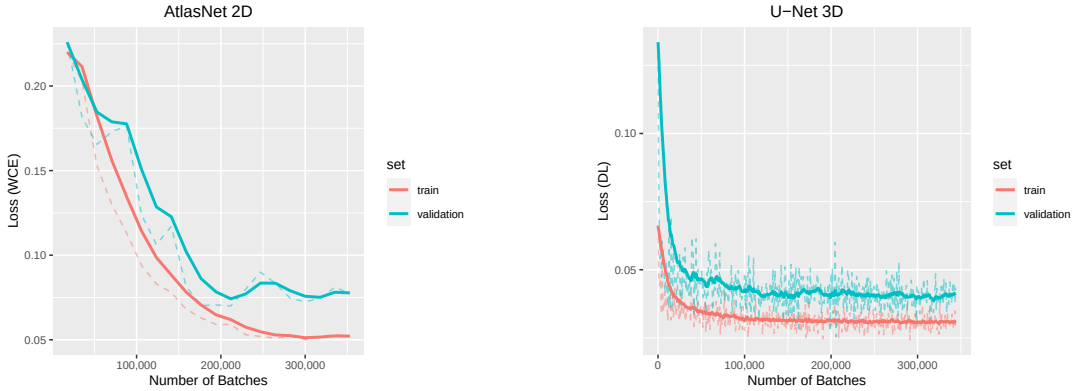


Figure 6. Training and validation curves for one template of AtlasNet and the 3D U-Net.

Statistical Analysis

The statistical analysis for the deep learning-based segmentation framework and the radiomics study was performed using Python 3.7, Scipy[35], Scikit-learn[36], TensorFlow[37] and Pyradiomics[34] libraries. The dice similarity score (DSC)[26] was calculated to assess the similarity between the 2 manual segmentations of each CT exam of the test dataset and between manual and automated segmentations. The DSC between manual segmentations served as reference to evaluate the similarity between the automated and the two manual segmentations. Moreover, the Hausdorff distance was also calculated to evaluate the quality of the automated segmentations in a similar manner. Disease extent was calculated by dividing the volume of diseased lung by the lung volume and expressed in percentage of the total lung volume. Disease extent measurement between manual segmentations and between automated and manual segmentations were compared using paired Student's t-tests.

For the stratification of the dataset into the different categories, classic machine learning metrics, namely balanced accuracy, weighted precision, and weighted specificity and sensitivity were used. Moreover, the correlations between each feature and the outcome was computing using a Pearson correlation over the entire dataset.

CT parameters between the 6 centers were compared using the analysis of variance, while patient characteristics between training/validation and test datasets were compared using chi-square and Student's t-tests.

	Training/ validation dataset			Testing dataset		
	Center A	Center B	Center C	Center D	Center E	Center F
CT equipment	Somatom AS+	Resolution HD	Aquilion Prime	Somatom Edge	Revolution HD	Aquilion Prime
Kilovoltage	100-120	120	100-120	100-120	120-140	100-120
DLP (mGy.cm)	109 ± 42 [44-256]	306 ± 104 [123-648]	102 ± 30 [43-189]	131 ± 44 [55-499]	177 ± 48 [43-276]	115 ± 26 [75 - 186]
CTDIvol (mGy)	3.2 ± 1.5 [1.2-11.9]	8.7 ± 2.8 [3.9-18.5]	2.7 ± 0.9 [1.0-5.3]	3.2 ± 0.9 [1.4-9.5]	5.5 ± 1.8 [1.2-12.3]	2.5 ± 0.6 [1.7-4.3]
Slice thickness	1mm	0.625mm	1mm	0.625mm	1mm	1mm
Convolution Kernel	i70	Lung	FC51-FC52	i50	Lung	FC51-FC52
Iterative reconstructions	SAFIRE 3	ASIR-v 80%	IDR 3D0.67	SAFIRE 4	ASIR-v 60%	IDR 3D

Table 1. Acquisition and reconstruction parameters. *Note.— For quantitative variables, data are mean $\pm$ standard deviation, and numbers in brackets are the range. CT = Computed Tomography ; CTDIvol = volume Computed Tomography Dose Index ; DLP = Dose Length Product * significant difference with* $p < 0.001$

References

1. Zu, Z. Y. *et al.* Coronavirus disease 2019 (covid-19): a perspective from china. *Radiology* 200490 (2020).

	Training/Validation Dataset (Centers A+B+C; N=50)	Test Dataset (Centers D+E+F; n=130)	p value
Age (y)	57 ± 17 [26-97]	59 ± 16 [17-95]	0.363
No. of Men	31(62)	87(67)	0.534
Disease extent*			
Manual	18.1 ± 14.9 [0.3-68.5]	19.5 ± 16.5 [1.1-75.7]	0.574
Automated	-	$19.9\% \pm 17.7$ [0.5-73.2]	-
DLP (mGy.cm)	180 ± 124 [43-527]	139 ± 49.0 [43-276]	0.026
CTDIvol (mGy)	4.9 ± 3.4 [1.0-13.0]	4.0 ± 1.9 [1.2-12.3]	0.064

Table 2. Patient characteristics in the datasets used for developing the segmentation tool. *Note.— For quantitative variables, data are mean ± standard deviation, and numbers in brackets are the range. For qualitative variables, data are numbers of patients, and numbers in parentheses are percentages. CTDIvol = volume Computed Tomography Dose Index; DLP = Dose Length Product *Percentage of lung volume on CT, calculated on the full volume for the training/validation dataset and 20 slices in the test dataset*

	Training/Validation Dataset (Centers A+B+D*; N=383)	Test Dataset (Centers C+E+F; n=95)	p value
Age (y)	63 ± 16 [24-98]	57 ± 15 [17-97]	0.001
No. of Men	255(67)	65(68)	0.732
Disease extent**	19.6 ± 17.0 [0.0-85.1]	22.5 ± 16.4 [1.1-64.7]	0.126
Short-term outcome			
Deceased	19(5)	8(8)	
Intubated	65(17)	18(19)	
Alive and Not Intubated	299(78)	69(73)	
LP (mGy.cm)	160 ± 97 [44-648]	146 ± 52.0 [43-276]	0.047
CTDIvol (mGy)	4.3 ± 2.8 [1.2-18.5]	4.1 ± 2.0 [1.0-12.3]	0.064

Table 3. Patient characteristics in the dataset used for the developed prognosis model using radiomics. *Note.— For quantitative variables, data are mean standard deviation, and numbers in brackets are the range. For qualitative variables, data are numbers of patients, and numbers in parentheses are percentages. CTDIvol = volume Computed Tomography Dose Index; DLP = Dose Length Product, ∗Enlarged by including all eligible patients over the study period, ∗∗Percentage of lung volume on the whole CT*

2. Bai, H. X. *et al.* Performance of radiologists in differentiating covid-19 from viral pneumonia on chest ct. *Radiology* 200823 (2020).
3. Bernheim, A. *et al.* Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection. *Radiology* 200463 (2020).
4. Li, K. *et al.* Ct image visual quantitative evaluation and clinical classification of coronavirus disease (covid-19). *Eur. Radiol.* 1–10 (2020).
5. Yuan, M., Yin, W., Tao, Z., Tan, W. & Hu, Y. Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in wuhan, china. *PLoS One* **15**, e0230548 (2020).
6. Truog, R. D., Mitchell, C. & Daley, G. Q. The toughest triage—allocating ventilators in a pandemic. *New Engl. J. Medicine* (2020).
7. White, D. B. & Lo, B. A framework for rationing ventilators and critical care beds during the covid-19 pandemic. *Jama* (2020).
8. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in china, 2019. *New Engl. J. Medicine* (2020).
9. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The Lancet* (2020).
10. Ai, T. *et al.* Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology* 200642 (2020).
11. Fang, Y. *et al.* Sensitivity of chest ct for covid-19: comparison to rt-pcr. *Radiology* 200432 (2020).

Classifier	Balanced Accuracy		Weighted Precision		Weighted Sensitivity		Weighted Specificity	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Nearest Neighbors	0.66 ±0.02	0.55 ±0.03	0.88 ±0.01	0.73 ±0.03	0.85 ±0.01	0.78 ±0.01	0.47 ±0.03	0.32 ±0.04
L-SVM*	0.71 ±0.02	0.67 ±0.04	0.80 ±0.02	0.77 ±0.03	0.68 ±0.02	0.67 ±0.04	0.74 ±0.03	0.67 ±0.06
P-SVM*	0.77 ±0.02	0.66 ±0.04	0.83 ±0.01	0.76 ±0.03	0.77 ±0.03	0.72 ±0.03	0.76 ±0.03	0.61 ±0.07
S-SVM	0.53 ±0.02	0.55 ±0.05	0.69 ±0.04	0.69 ±0.04	0.49 ±0.08	0.51 ±0.1	0.57 ±0.09	0.58 ±0.11
RBF-SVM*	0.74 ±0.02	0.67 ±0.06	0.82 ±0.01	0.77 ±0.04	0.66 ±0.02	0.63 ±0.06	0.81 ±0.02	0.71 ±0.08
Gaussian Process	0.63 ±0.03	0.59 ±0.04	0.82 ±0.02	0.77 ±0.04	0.83 ±0.01	0.80 ±0.02	0.43 ±0.04	0.38 ±0.05
Decision Tree*	0.78 ±0.02	0.68 ±0.04	0.85 ±0.01	0.78 ±0.03	0.69 ±0.03	0.64 ±0.05	0.87 ±0.03	0.72 ±0.08
Random Forest*	0.78 ±0.02	0.64 ±0.06	0.84 ±0.02	0.75 ±0.04	0.74 ±0.03	0.65 ±0.07	0.82 ±0.04	0.63 ±0.09
Multi-Layer Perceptron	0.83 ±0.04	0.58 ±0.03	0.90 ±0.02	0.72 ±0.02	0.9 ±0.02	0.73 ±0.03	0.77 ±0.06	0.43 ±0.05
AdaBoost*	0.8 ±0.03	0.63 ±0.03	0.86 ±0.01	0.75 ±0.03	0.76 ±0.07	0.66 ±0.07	0.83 ±0.08	0.6 ±0.11
Gaussian Naive Bayes*	0.66 ±0.02	0.63 ±0.05	0.76 ±0.01	0.74 ±0.03	0.74 ±0.01	0.72 ±0.04	0.57 ±0.03	0.53 ±0.07
Bernouilli Naive Bayes	0.51 ±0.01	0.50 ±0.01	0.74 ±0.05	0.63 ±0.05	0.78 ±0.0	0.77 0.01	0.24 ±0.01	0.22 ±0.02
QDA	0.72 ±0.01	0.60 ±0.04	0.82 ±0.01	0.73 ±0.03	0.82 ±0.01	0.75 ±0.03	0.62 ±0.02	0.45 ±0.07

Table 4. Performances of the 13 evaluated classifiers. *Note.— Asterixis indicates the 7 classifiers reporting a balanced accuracy higher than 0.60 and that were finally selected, L-SVM = Support Vector Machine with a linear kernel; L-SVM = Support Vector Machine with a polynomial kernel, L-SVM = Support Vector Machine with a sigmoid kernel, RBF-SVM = Support Vector Machine with a Radial Basis Function, QDA = Quadratic Discriminant Analysis*

12. Xie, X. *et al.* Chest ct for typical 2019-ncov pneumonia: relationship to negative rt-pcr testing. *Radiology* 200343 (2020).

13. Chassagnon, G., Vakalopoulou, M., Paragios, N. & Revel, M.-P. Artificial intelligence applications for thoracic imaging. *Eur. J. Radiol.* **123**, 108774 (2020).

14. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017).

15. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. medicine* **25**, 954–961 (2019).

16. Wang, L. & Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images. *arXiv preprint arXiv:2003.09871* (2020).

17. Li, L. *et al.* Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology* 200905 (2020).

18. Chaganti, S. *et al.* Quantification of tomographic patterns associated with covid-19 from chest ct. *arXiv preprint arXiv:2004.01279* (2020).

19. Huang, L. *et al.* Serial quantitative chest ct assessment of covid-19: Deep-learning approach. *Radiol. Cardiothorac. Imaging* **2**, e200075 (2020).

20. Jacob, J. *et al.* Mortality prediction in ipf: evaluation of automated computer tomographic analysis with conventional severity measures. *Eur. Respir. J.* ERJ–01011 (2016).

21. Humphries, S. M. *et al.* Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. *Radiology* **285**, 270–278 (2017).

Features		Correlation					
		Severe vs Non Severe		Intubated vs Diseased		Hierarchical Classifier	
Age		-0.0681		0.6745		0.1853	
Gender		-0.1517		0.0211		0.1443	
Disease Extent		-0.3809		-0.0280		0.3491	
Heart	non-uniformity on the GLSZM	-0.1953		-0.0935		0.1537	
		Left	Right	Left	Right	Left	Right
Lung	Skewness	0.3324	0.3675	0.1018	-0.0229	-0.2967	-0.3447
	90th Percentile	-0.2808	-0.3221	-0.0576	0.0035	0.2526	0.3119
Disease	Maximum attenuation	-0.0867	-0.1978	0.0885	0.0035	0.0958	0.1846
	Surface	-0.3382	-0.3289	-0.0174	0.0381	0.3034	0.3112
	Maximum 2D diameter per slice	-0.2304	-0.2236	0.1238	0.1010	0.2288	0.2183
	Volume	-0.3592	-0.3971	-0.0489	0.0840	0.3241	0.3863
	Non-uniformity on the GLSZM	-0.2944	-0.3448	-0.1379	-0.0805	0.2375	0.3004
	Non-uniformity on the GLRLM	-0.3280	-0.3372	-0.0219	0.0829	0.3000	0.3319

Table 5. Correlation between outcome and the 12 selected features. *Note.—Note.— GLSZM =Gray level Size Zone matrix , GLRLM = Gray level Run Length matrix*

22. Vakalopoulou, M. *et al.* Atlasnet: Multi-atlas non-linear deep networks for medical image segmentation. In Frangi, A. F., Schnabel, J. A., Davatzikos, C., Alberola-López, C. & Fichtinger, G. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 658–666 (Springer International Publishing, Cham, 2018).

23. Anthimopoulos, M. *et al.* Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. *IEEE journal biomedical health informatics* **23**, 714–722 (2018).

24. Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis machine intelligence* **39**, 2481–2495 (2017).

25. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*,

Classifier	Balanced Accuracy		Weighted Precision		Weighted Sensitivity		Weighted Specificity	
	Training	Test	Training	Test	Training	Test	Training	Test
L-SVM	0.71	0.7	0.8	0.76	0.68	0.67	0.75	0.73
P-SVM	0.76	0.67	0.82	0.73	0.75	0.69	0.76	0.65
RBF-SVM	0.72	0.71	0.81	0.79	0.65	0.63	0.79	0.79
Decision Tree	0.76	0.65	0.85	0.74	0.65	0.56	0.88	0.74
Random Forest	0.76	0.7	0.82	0.76	0.71	0.65	0.8	0.75
AdaBoost	0.77	0.67	0.83	0.74	0.71	0.62	0.82	0.71
Gaussian Naive Bayes	0.65	0.73	0.76	0.77	0.73	0.76	0.57	0.69
Ensemble Classifier	0.75	0.74	0.82	0.79	0.69	0.69	0.81	0.79

Table 6. Performances of each of the 7 individual classifiers and of the ensemble classifier to differentiate between patient with severe and non-severe short-term outcome. *Note.— L-SVM = Support Vector Machine with a linear kernel; L-SVM = Support Vector Machine with a polynomial kernel, RBF-SVM = Support Vector Machine with a Radial Basis Function kernel*

Classifier	Balanced Accuracy		Weighted Precision		Weighted Sensitivity		Weighted Specificity	
	Training	Test	Training	Test	Training	Test	Training	Test
L-SVM	0.84	0.78	0.87	0.84	0.81	0.85	0.87	0.72
P-SVM	0.95	0.81	0.96	0.9	0.95	0.88	0.95	0.74
RBF-SVM	0.9	0.62	0.92	0.83	0.90	0.77	0.9	0.48
Decision Tree	0.97	0.75	0.98	0.87	0.98	0.85	0.96	0.65
Random Forest	0.94	0.75	0.96	0.87	0.96	0.85	0.92	0.65
AdaBoost	1	0.62	1	0.83	1	0.77	1	0.48
Gaussian Naive Bayes	0.82	0.75	0.86	0.87	0.83	0.85	0.8	0.65
Ensemble Classifier	0.96	0.81	0.97	0.9	0.96	0.88	0.95	0.74

Table 7. Performances of each of the 7 individual classifiers and of the ensemble classifier to differentiate between intubated and deceased patients. *Note.— L-SVM = Support Vector Machine with a linear kernel; L-SVM = Support Vector Machine with a polynomial kernel, RBF-SVM = Support Vector Machine with a Radial Basis Function kernel*

424–432 (Springer, 2016).

26. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).

27. Sun, R. *et al.* A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet Oncol.* **19**, 1180–1191 (2018).

28. Wang, W. *et al.* Detection of sars-cov-2 in different types of clinical specimens. *Jama* (2020).

29. CDC. Triage of suspected covid-19 patients in non-us healthcare settings. centers for disease control and prevention. *https://www.cdc.gov/coronavirus/2019-ncov/hcp/non-us-settings/sop-triage-prevent-transmission.html* (2020).

30. Tang, N., Li, D., Wang, X. & Sun, Z. Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. *J. Thromb. Haemostasis* (2020).

31. Onder, G., Rezza, G. & Brusaferro, S. Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy. *Jama* (2020).

32. Guo, W. *et al.* Diabetes is a risk factor for the progression and prognosis of covid-19. *Diabetes/Metabolism Res. Rev.* (2020).

33. Ferrante, E., Dokania, P. K., Marini, R. & Paragios, N. Deformable registration through learning of context-specific metric aggregation. In *International Workshop on Machine Learning in Medical Imaging*, 256–265 (Springer, 2017).

34. Van Griethuysen, J. J. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer research* **77**, e104–e107 (2017).

35. Virtanen, P. *et al.* Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. methods* 1–12 (2020).

36. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine learning research* **12**, 2825–2830 (2011).

37. Abadi, M. *et al.* Tensorflow: Large-scale machine learning on heterogeneous systems. (2015).

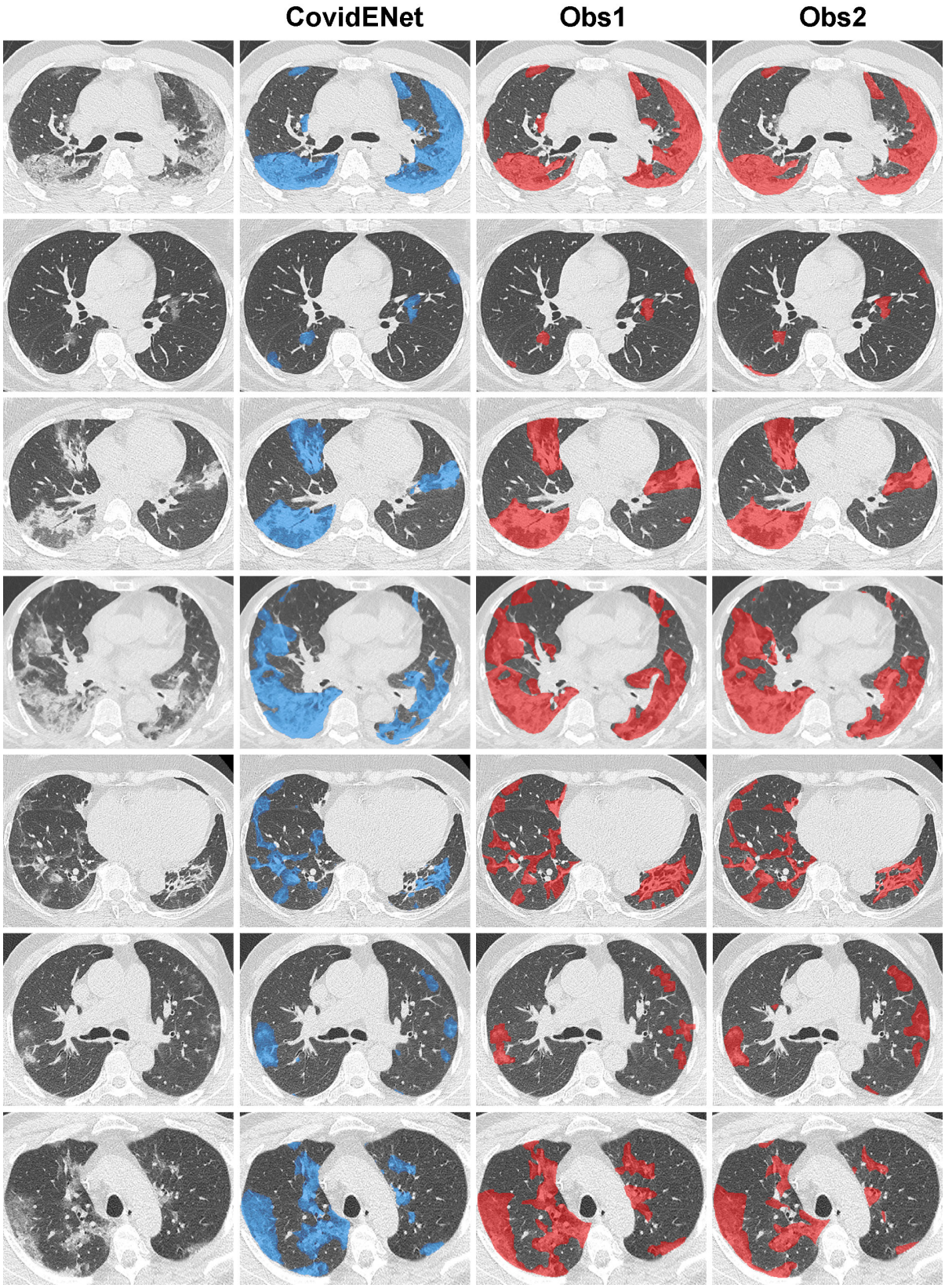


Figure 7. Some additional qualitative analysis for the comparison between automated and manual segmentations. Delineation of the diseased areas on chest CT in different slices of COVID-19 patients: From left to right: Input, AI-segmentation, expert I-segmentation, expert II-segmentation

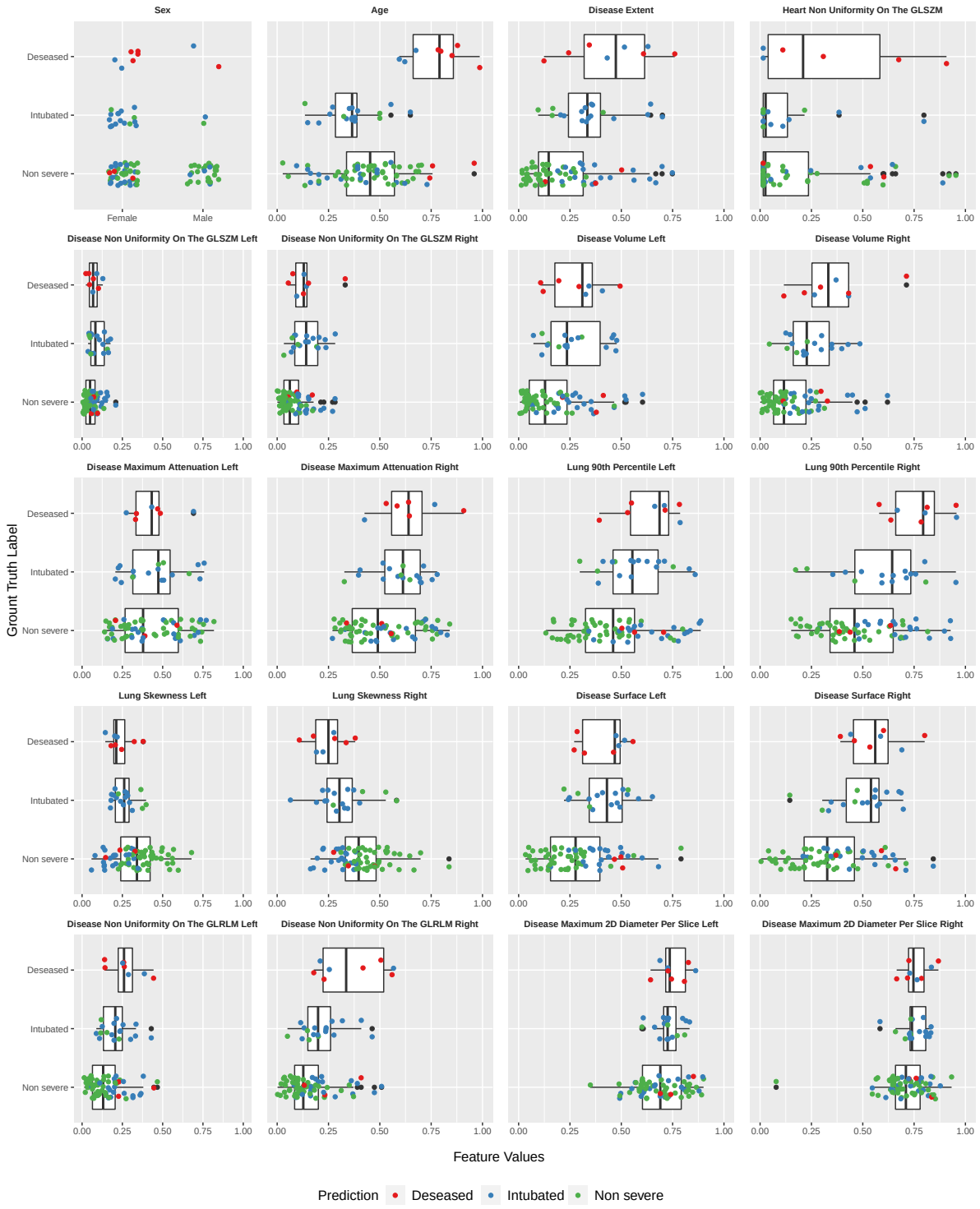


Figure 8. Boxplots of the selected features and their association with the predicted outcomes and ground truth labels.