Database collection

This database contains data gathered across six different hospitals in northern Italy during the first outbreak of SARS-CoV-2 (March 2020). All included subjects were confirmed with a diagnosis of COVID-19 following hospitalization. Disease outcome was updated at a later stage and is here reported as either severe (if patient required forced mechanical ventilation or died) or mild (all other outcomes).

During triage, a set of clinical tests were performed generating several clinical parameters, 16 of which were deemed relevant for outcome prediction and included in the dataset. The following table reports name, a brief description and average values over the training set of the collected items. **Not all items are available for all subjects.**

Features (numerical variables)	Description	Mean (train set)	Std (train set)
Age	Patient's age (years)	64.44	15.05
Body Temperature (°C)	Patient temperature at admission (in °C)	37.60	0.97
WBC	White blood cells count (10^9/L)	7.06	3.53
CRP	C-reactive protein concentration (mg/dL)	40.68	66.93
Fibrinogen	Fibrinogen concentration in blood (mg/dL)	602.52	158.36
LDH	Lactate dehydrogenase concentration in blood (U/L)	368.17	235.22
D-dimer	D-dimer amount in blood	2.5e+03	6743.13
O_2	Oxygen percentage in blood	92.54	7.00
PaO_2	Partial pressure of oxygen in arterial blood (mmHg)	72.09	26.11
SaO_2	Arterial oxygen saturation (%)	91.92	8.24
pH	Blood pH	7.45	0.06

Features (binary variables)	Description	Mean (train set)
Sex	Patient's sex (0 – male, 1 – female)	0.34
Cough	Cough (0 – no, 1 – yes)	0.51
Dyspnea (DifficultyInBreathing)	Patient had intense tightening in the chest, air hunger, difficulty breathing, breathlessness or feeling of suffocation (0 – no, 1 – yes)	0.50
Cardiovascular Disease	Patient had cardiovascular disease (0 – no, 1 – yes)	0.28
Respiratory Failure	Patient had respiratory failure (0 – no, 1 – yes)	0.02

For each patient, a single chest X-ray is reported, also collected on first day of admission. X-ray scans often occurred in emergency conditions, therefore both image quality and subject position are highly variable. Furthermore, some images were collected on digital support, while others are the result of digitalization of film images. While all images are 16-bit png files, intensity levels for most of them do not cover the entire possible range of values, given their different origins.

For the purpose of this challenge, data from 5 of the 6 centers constitute the training set and will be provided as collected (raw, 863 subjects). Data from the last center (120 subjects) is used as testing set to rank the solutions. In addition, data in the test set IS NOT provided entirely: 20% of the x-ray images, as well as 30% of the available entries for each clinical parameter have been removed. Disease severity is provided in all instances for the training set and never for the test set.

A more detailed description of the dataset, as well as early attempts to solve a problem similar to Task 2 (see below) can be found in [1].

[1] Soda, Paolo, et al. "AIforCOVID: predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study." arXiv preprint arXiv:2012.06531 (2020).

Tasks description

This challenge is divided in two separate sub-tasks, which will be evaluated separately.

1 Task 1: Clinical data imputation

The first task consists in estimating the value of missing entries in the clinical information of the test set. The metric used to rank proposed solutions is the mean squared estimation error, with two modifications. In order to compensate for different ranges in the various fields, data will be normalized by the standard deviation of available values of the training set before score computation. Furthermore, to limit the impact of possible outliers, for each single entry, error will saturate at 1 SD. We will therefore consider the following metric for task 1:

$$m_1^i = min\left(\left(\frac{X_i - \widehat{X_\imath}}{\sigma_{x_j}}\right)^2, 1\right)$$

where X_i and $\widehat{X_\imath}$ are, respectively, the real and estimated values for entry i. σ_{x_j} is the standard deviation of field j computed on the training set (values reported in the table above) for numerical variables, while it is set to 1 for binary ones.

Participants are expected to provide a reconstruction value for each missing entry, but score will be computed only on those that have been artificially removed.

2 Task 2: Disease severity prediction

The second task consists in the classification of subjects according to severity in two classes: 'MILD' vs 'SEVERE'. In this case, the metric used for ranking is simply the accuracy value of the proposed classification on the test set. Please note that proportion of classes may vary between training and test sets.

$$m_2^p = 1\ if\ c_p = \widehat{c_p}, else\ 0$$

where c_p and $\widehat{c_p}$ are, respectively, the ground truth and predicted severity class for patient p in the test set.

3 Use of other data sources

The use of *publicly available* data sources and pre-trained models is allowed during the challenge. When you start using another source of data, you are required to share this information with fellow participants in the Slack channel '#data-sources'.

4 Submission of results

To submit your results, each team should use the CSV file *testSet.txt* containing the clinical data of the test set as a template, with the following modifications:

- *Rename the file* with your team's name
- Replace the *missing fields 'NaN'* with the imputed values for the clinical parameters from task 1 (integer or scalar value)
- Replace the *'<undefined>' Prognosis field* with your class prediction from task 2 (either 'MILD' or 'SEVERE')

5 Ranking of teams

For each sub-challenge, the teams will be ranked using the *significance ranking* method proposed in [2]. This method employs pairwise significance tests that assess significant differences in metric values between algorithms, and ranks the algorithms according to the number of algorithms to which they are significantly superior to.

In *case of ties* for the first place (and only in this case), the winning algorithm will be chosen based on the average metrics. For task 1, the team with the smallest mean imputation score ($S_1 = \frac{1}{N}\sum_{i=1}^{N} m_1^i$, where N is the number of artificially removed entries in the test set) would then be declared as winner. For task 2, the team with the largest prediction accuracy ($S_2 = \frac{1}{P}\sum_{p=1}^{P} m_2^p$, where P is the number of patients in the test set) would be declared as winner.

[2] Wiesenfarth, Manuel, et al. "Methods and open-source toolkit for analyzing and visualizing challenge results." *Scientific Reports* 11.1 (2021): 1-15.

Additional information

Beside the clinical information and original chest x-rays data ('rawImg' folder) two additional image archives are made available: a pre-processed version of x-ray images ('normalizedImg' folder) and a collection of radiomic features ('Radiomics' folder). X-ray images have been resized to the constant shape of 1200 x 1200 pixels with a bicubic interpolation. Furthermore, intensity ranges have been normalized: pixels with values below the first (above the 99th) percentile have been set to the first (99th) percentile; finally, images underwent histogram equalization so that the histogram of the output image approximately matches a "flat" histogram.

Radiomic features have been computed on 7 x 7 patches, with step 1 in each direction. A symmetric padding of the images was adopted to compute values close to the borders, therefore radiomic feature maps have the same size as the pre-processed images (1200 x 1200 pixels). The exact description of the equations used to compute each radiomics feature is provided in the attached pdf file, an extract of the publication in which the dataset was originally described.

Please note that neither pre-processed images nor radiomics features are in any way required for either of the tasks proposed above: they have been included as some participants might attempt to devise a solution that makes use of them. The time required to compute radiomic features of the presented dataset is not compatible with the duration of the challenge: as a reference, it took 72 hours to compute the feature maps of the 1200 x 1200 pre-processed images.