Covid CXR Hackathon

Alberto Presta, Carlo Alberto Barbano, Enzo Tartaglione, Marco Grangetto

The aim of our algorithm is to build a process that automatically classifies the severity of the prognosis, either it being severe or mild, in patients affected by COVID-19. We have designed a classifier using clinical data and radiographic findings, extracted from the CXRs. We leveraged on a method mimicking a radiographer's diagnostic process: we extract radiological observations from the CXR images, and jointly with other available metadata, we trained a decision tree model to elaborate the diagnosis. We used explainable models for classifying the severity of the infection, as soon our approach is able to provide a clinical motivation for the diagnosis. We choose this way because other existing techniques used to interpret the result of a Neural network, like Gradcam, suffer from high variability and thus they are not truly reliable.

I. Data Preprocessing

A. CXR Images

For each CXR image, we perform the following steps.

1) We remove the padding around the image. In particular, we delete an edge if it was composed only by pixels with values below 5% or above 95% of the maximum encoded value.
2) We remove some superimposed writings. We exploit the built-in function `inpaint` from the OpenCV package.
3) We convert the image to 8-bits per pixel, and we standardize it.
4) We transform the image to `MONOCHROME2`. We take a small central square in the central portion of the image, which is empirically supposed to be bright given the average patients pose: if the average of the pixels is dark, then we invert the pixels.
5) We resize the images to 512×512, and then we take a central crop of size 448×448.

B. Clinical Data

Among the metadata provided within the dataset, there is a consistent percentage of missing values. These have been filled, at training time, in the following way:

- the mean value has been used for continuous metadata;
- the mode has been used for binary metadata;
- the round integer approximation of the mean has been used for discrete metadata.

We have not used all the variables in our learning process, but we have made a selection based on two factors:

1) the Pearson correlation coefficient with respect to the specific target, calculated in the training set, at least 0.09 (empirically determined);
2) the percentage of missing values, not exceeding 25%.

Under these restrictions, for the prognosis severity classification task we will use the following 9 features:

- *Age*,
- *LDH*,
- *DifficultyInBreathing*,
- *WBC*,
- *CPR*,
- *HighBloodPressure*,
- *Diabetes*,
- *Glucose*,
- *BPCO*.

It has been observed that the level of oxygen in the blood is fundamental towards a correct diagnosis; in fact, in absolute value it has the highest Pearson correlation coefficient (0.39) with respect to the Prognosis. Unfortunately, in the train set this value is present only in about 67% of the patient, and it is missing in the test set. Hence, a prediction of such a value will be performed, using either the features extracted from the CXR image and the following 6 metadata:

1) *DifficultyInBreathing*,
2) *PaO2*,
3) *Age*,
4) *CRP*,
5) *Glucose*.

II. The algorithm

We divide the presentation of the model in three subsections, namely:

1) a brief description of the backbone;
2) the estimation of a new binary variable, called *Oxygen percentage*, which tells us if a patient has a level of saturation below or above 92% which represents the average value over the training set;
3) the actual binary classification that predicts, using both clinical data, the new synthetic variable, and features extracted by images, the severity of the COVID-19 infection.

A high-level blueprint of the entire process is pictured in figure 1.

A. Radiological findings extractor

For what it concerns the extraction of features form the CXR image, the same approach as in [1] has been used. In particular, from the CXR image we aim at identifying the presence/absence of 14 specific radiological observations, from the radiographic image. The model's encoder is the same as DenseNet-121's. This model has been trained on the CheXpert dataset [2], using SGD with learning rate 0.01 and weight decay 0.0001, optimizing binary cross-entropy as loss function.

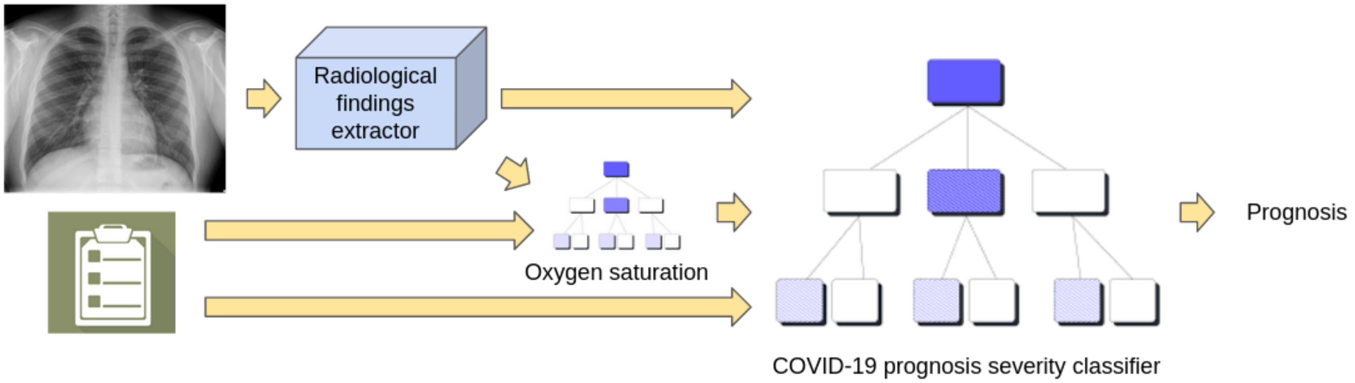
Radiological findings extractor
Oxygen saturation
Prognosis
COVID-19 prognosis severity classifier

Fig. 1: Blueprint of the algorithm for COVID-19 prognosis severity classification. The inputs are respectively the image and the clinical data.

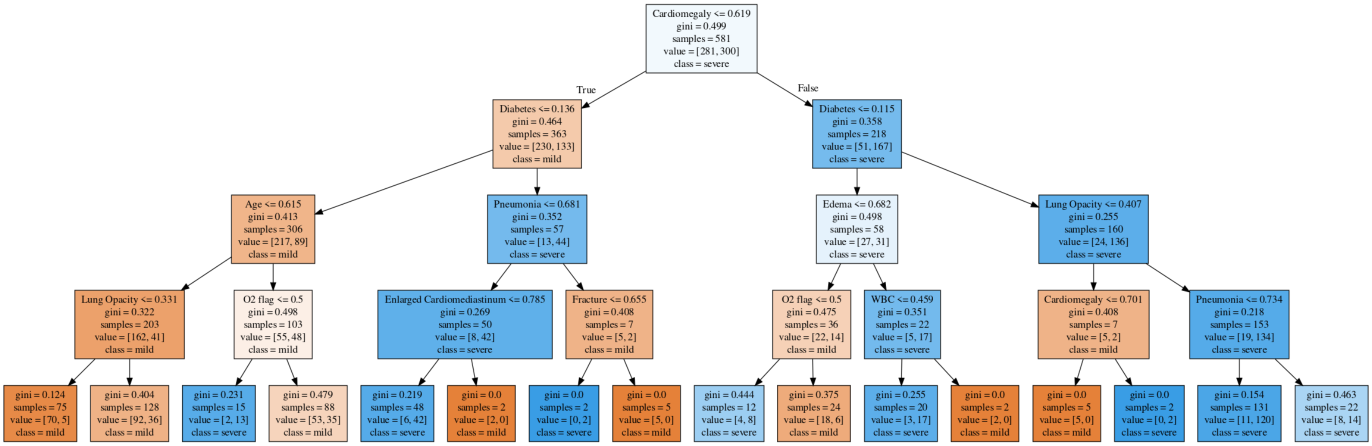

Fig. 2: The obtained decision tree to classify the COVID-19 severity.

B. *Oxygen saturation*

In order to estimate this important feature, not present in either CheXpert or the challenge's test set -despite very well correlating with the target, we decide to train a binary classification tree able to predict binary variable that tells us if the oxygen saturation level is above 92% or not (HIGH or LOW). We choose this threshold because it represents the average of this field in the train set. As training data we concatenated the 6 features mentioned in I-B, and the the 14-dimensional vector extracted from the radiological findings extractor, obtaining thus a 20-dimensional vector. For what concerns the training phase, we have performed a 5-fold cross validation setting to evaluate the model. In order to find the configuration that maximizes the overall accuracy, we have applied a grid research across the possible maximum depth (4 to 10) of the tree and the splitting criterion (either Gini or Entropy).

C. *COVID-19 prognosis severity classifier*

To perform binary classification over the COVID-19 Prognosis we have trained a binary classification tree, and we have used as training data the concatenation of the 14-dimensional vector extracted with the radiological findings extractor, the 9 variables mentioned in I-B, and the Oxygen saturation level, obtaining thus a 24-dimensional vector for each patient. We designated 5-fold cross validation setting to evaluate the model, and we applied grid research across the criterion to perform the splits (either Gini or Entropy) and the maximum level of the tree (from 4 to 10) to find the optimal configuration. The resulting decision tree, used for the diagnosis, is shown in figure 2.

References

[1] Carlo Alberto Barbano and Enzo Tartaglione and Claudio Berzovini and Marco Calandri and Marco Grangetto, *A two-step explainable approach for COVID-19 computer-aided diagnosis from chest x-ray images*, 2021

[2] Jeremy A. Irvin and Pranav Rajpurkar and Michael Ko and Yifan Yu and Silviana Ciurea-Ilcus and Chris Chute and Henrik Marklund and Behzad Haghgoo and Robyn L. Ball and Katie S. Shpanskaya and Jayne Seekins and David A. Mong and Safwan S. Halabi and Jesse K. Sandberg and Ricky H Jones and David B. Larson and C. Langlotz and Bhavik N. Patel and Matthew P. Lungren and A. Ng, *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*, AAAI, 2019