Hackathon Covid

RICCARDO AGAZZOTTI

244836@studenti.unimore.it

FEDERICO COCCHI

289842@studenti.unimore.it

February 2022

1 Purposes

The purpose of this work is to define an explainable way to define if a COVID patient health status is "MILD" or "SEVERE" using an algorithm which elaborates this dataset [8.1]. Each one of the following section explains one step of our classification pipeline.

2 Preprocessing

Analyzing the provided dataset [8.1], one can easily see that some images are dark or with a very low gradient, thus to make them work better for segmentation and classification we processed them using classical image processing techniques as linear luminance adjustment, normalization and contrast fixing. This step was also essential to make our images more related to the ones in the Montgomery dataset [8.2], which is the one used for the segmentation network training. Going through these simple normalization steps the performance of the whole network increases significantly.
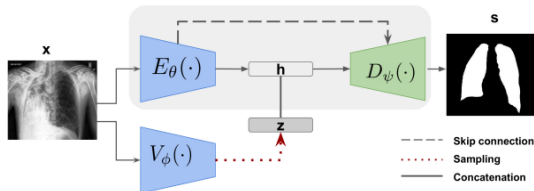
3 Segmentation

Figure 1: Segmentation network

After the image preprocessing our goal is to extract the lungs since is the area we have to focus on for further information extraction about COVID disease. A segmentation technique has been used for pursuing this goal and cleaning the semantic noise of the non-lung area. In particular the implemented architecture uses a VAE and a UNET [3]. The VAE is used for performing a cross-domain mapping between the input X and the segmentation target S. The architecture shows two encoders. One related to the UNET [2] and the other one for the VAE. The latent vector (z), which will be forwarded in the decoder, is a concatenation of the output of the Unet Encoder and a sampled vector from the probability distributions given by the VAE Encoder. The reconstruction loss used is the binary cross entropy loss since the masks on which the network has been trained are binary.

4 Classification

After the preprocessing part with also data augmentation 8, we cleaned the semantic noise provided from the non-lungs area. Since we had to extract class information from the images we trained a resnet101 to classify our CXR images as "MILD" or "SEVERE", thus generating a column data to add to our 39 values dataset. The produced dataset, with image information, is classified using a decision tree (Entropy splitting) as main technique. In the results section [7] the reached performance are highlighted.

5 Explainability

One of the main challenges that we had to face in this work are related to explainability. To pursue this goal in the most complete way possible, the problem has been split in two parts:

- Explainability for image classification

- Explainability for data classification

1

In terms of image classification we have trained a RESNET101 to get the final label and to explain the behavior of the resnet the last convolutional layer activation map has been extracted and used for generating an heatmap which tells us where, on the lungs area, the network has activated to provide the final label.

In terms of data classification we used decision tree as main technique, using all clinic data plus the information from resnet classification (binary classification on images). So, merging the two type of informations and computing on it the evaluation with a decision tree.

Thus, by plotting the decisions taken at each node, as in 7. So, we completely define how our final label has been reached in a explainable way,

6 Pipeline and Examples

In this image we can see the various steps of the pipeline of our model, considering the part of classification and part of heatmap. In both of them we have used methodologies to help human to undersand what happens in the network.
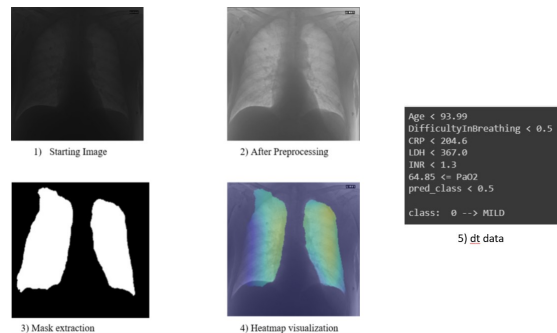

Figure 2: Pipeline

Figure 3: Decision tree outcome

7 Results

To create this table we split a small portion of the training data to evaluate our model and to verify the improvement of the accuracy merging clinical data and classified images. For the second accuracy metrics we use the formula of the algorithm evaluation in the FAQ page.

Model	accuracy	acc_hac
dummy classifier	0.46	0.45
decision tree	0.67	0.65
decision tree with images	0.92	0.91
random forest	0.75	0.75
random forest with images	0.97	0.96

Table 1: Comparison of results on evaluation

8 Datasets

In this section is provided a breif description of the used datasets to train the previously described networks (segmentation, classification).

For the project we have not a big quantity of data, so we used techniques of data augmentation to resolve the problem and to achieve a more general behaviour of the model.

8.1 Hackton Dataset

This dataset provides for the train part 1100 elements. Each element is composed of a CXR image and 39 more vital signs to define in the most accurate way the global health of a COVID patient. Each data has it own label "MILD" or "SEVERE" [1].

8.2 Montgomery Dataset

Montgomery dataset has been used in our work to train the segmentation network. In particular the dataset is composed of roughly 500 CXR (4892x4020) images and for each image the lungs segmentation mask is provided.

References

[1] Aiforcovid: Predicting the clinical outcomes in patients with covid-19 applying ai to chest-x-rays. an italian multicentre study, 2021.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[3] Raghavendra Selvan, Erik B. Dam, Nicki S. Detlefsen, Sofus Rischel, Kaining Sheng, Mads Nielsen, and Akshay Pai. Lung segmentation from chest x-rays using variational data imputation, 2020.