Covid-19 Detection from Lung CT images

Mattia Rigiroli - 819978
Simone Magnaghi – 813070
Enrico Conte – 852679

Table of Contents

01 Problem description

02 Dataset

03 Pipeline

04 Results

05 Conclusions

1. Problem Description

The goal of this research is to train a **classifier** to recognize Covid-19 positive patients from their **CT lungs scans** in order to **support** the physician's **decision process** with a quantitative approach.

2. Dataset

Negatives

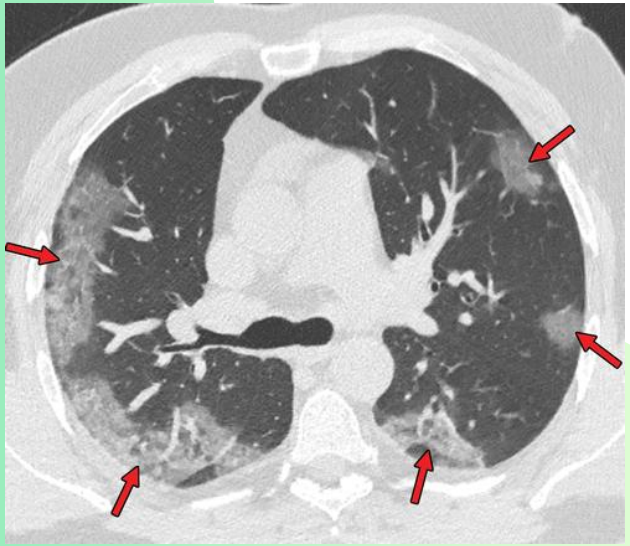
118 Patients
8GB of DICOM Images

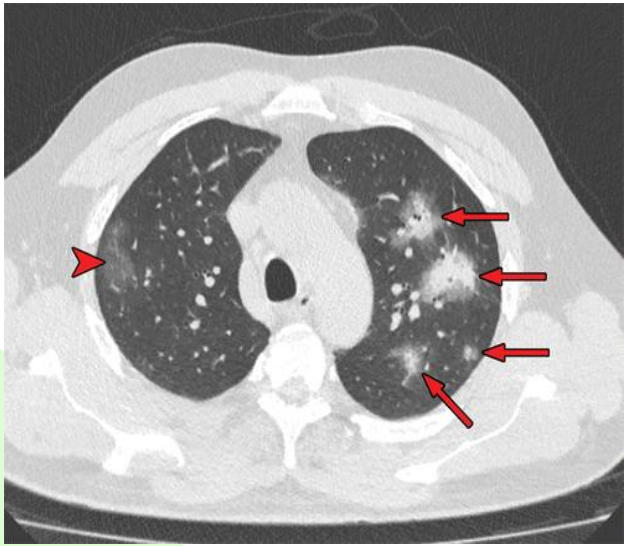
Source:
https://wiki.cancerimagingarchive.net/

Positives

111 Patients
11 GB of DICOM Images

Negatives





Positives

Chest CT in COVID-19: What the Radiologist Needs to Know
Thomas C. Kwee and Robert M. Kwee
RadioGraphics 2020 40:7, 1848-1865
https://pubs.rsna.org/doi/pdf/10.1148/rg.2020200159

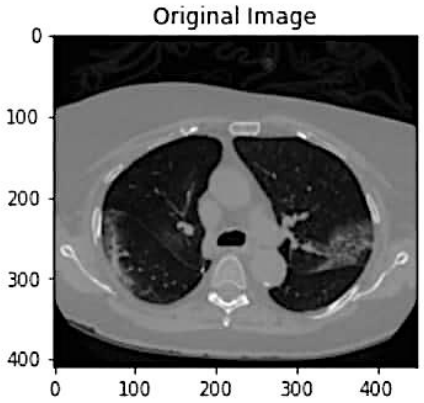
3. Pipeline

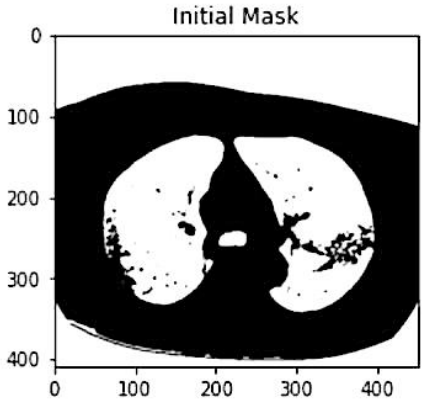
Per Patient

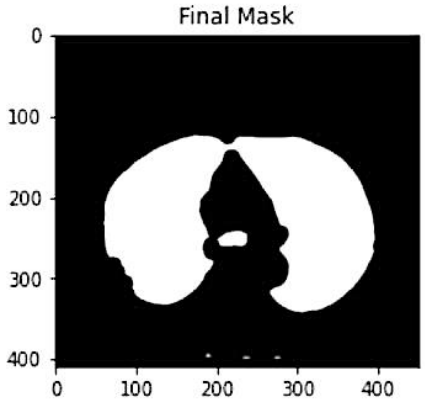
- Slices Selection x3
- Mask Generation
- Mask Fill
- Histogram Equalization and Filtering
- Haralick Features Extraction

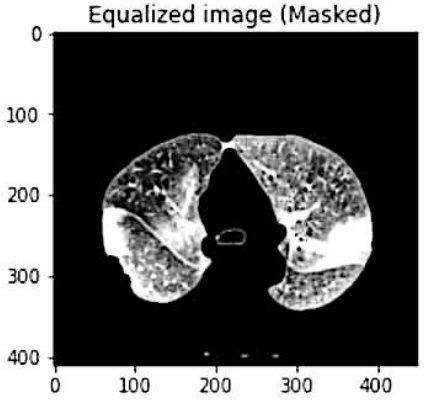
Whole Dataset

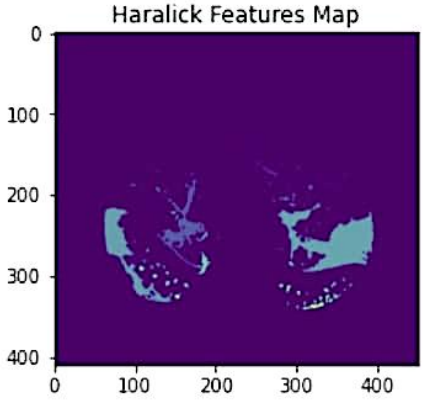
- Feature Reshaping
- Feature Reduction through PCA
- Feeding to the Classifier

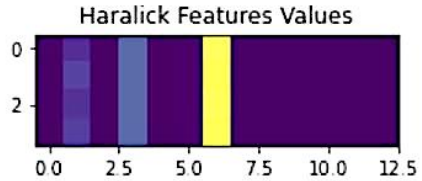






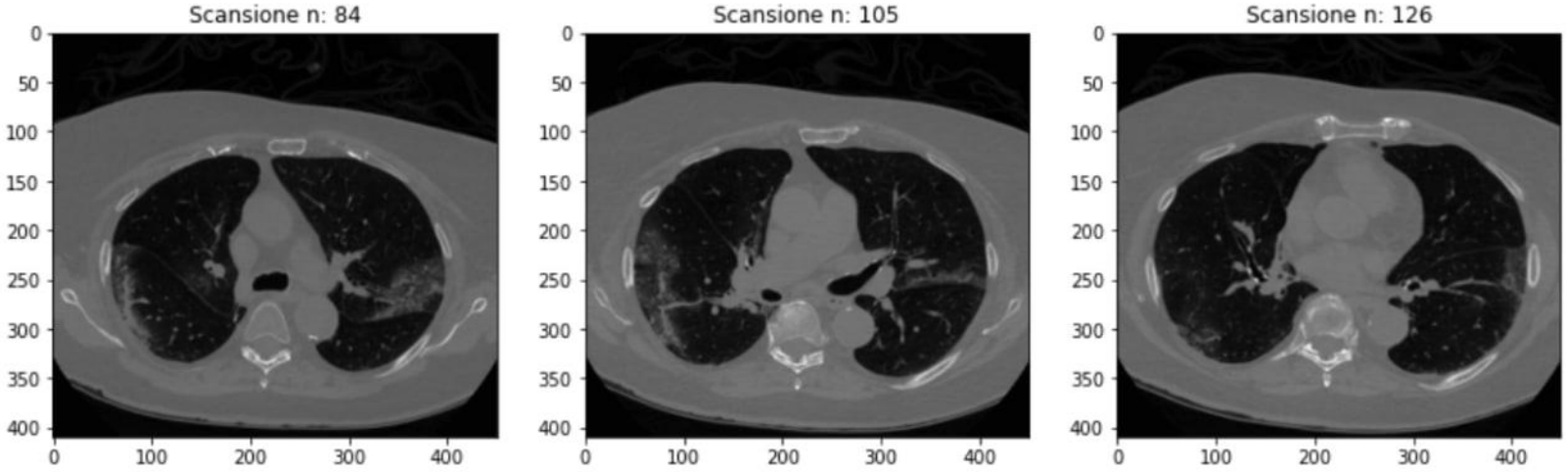






Slices Selection

For each patient exam we select **three scans** at the **33rd**, **41st** and **49th percentile** of the total number of available scans: the idea is to **isolate** the **most representative slices.**

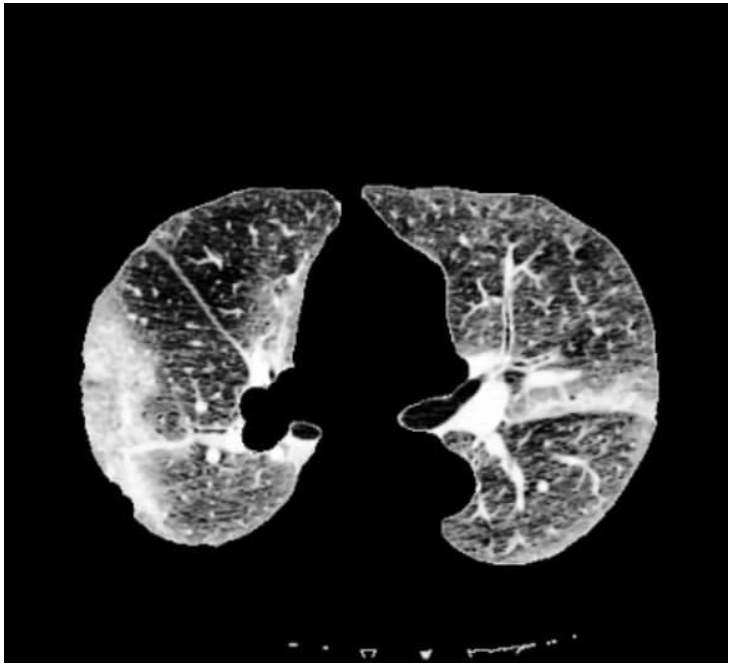


Mask Generation and Fill

First the mask is generated via **Kmeans Clustering** (**K=2**).

Afterwards we **erase** the borders, **erode** the pixels and **fill** the remaining holes to obtain a more robust representation of the lung.

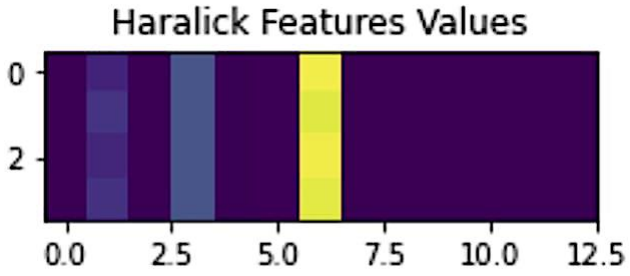
Histogram Equalization and Filtering



In order to highlight the regions of interest, we performed the **equalization** of the **Histogram** of the masked image.

Then, to reduce the overall noise content of the image we applied a **Wiener Filter.**

Haralick Features Extraction



We extracted the **Haralick Features** from the equalized and filtered masked image.

The idea is to extract the **textural properties** of the regions of interest and feed them into the classifier.

The output is a 4x13 matrix, where 4 is the number of directions and 13 is the number of features.

Feature Reduction through PCA and Modeling

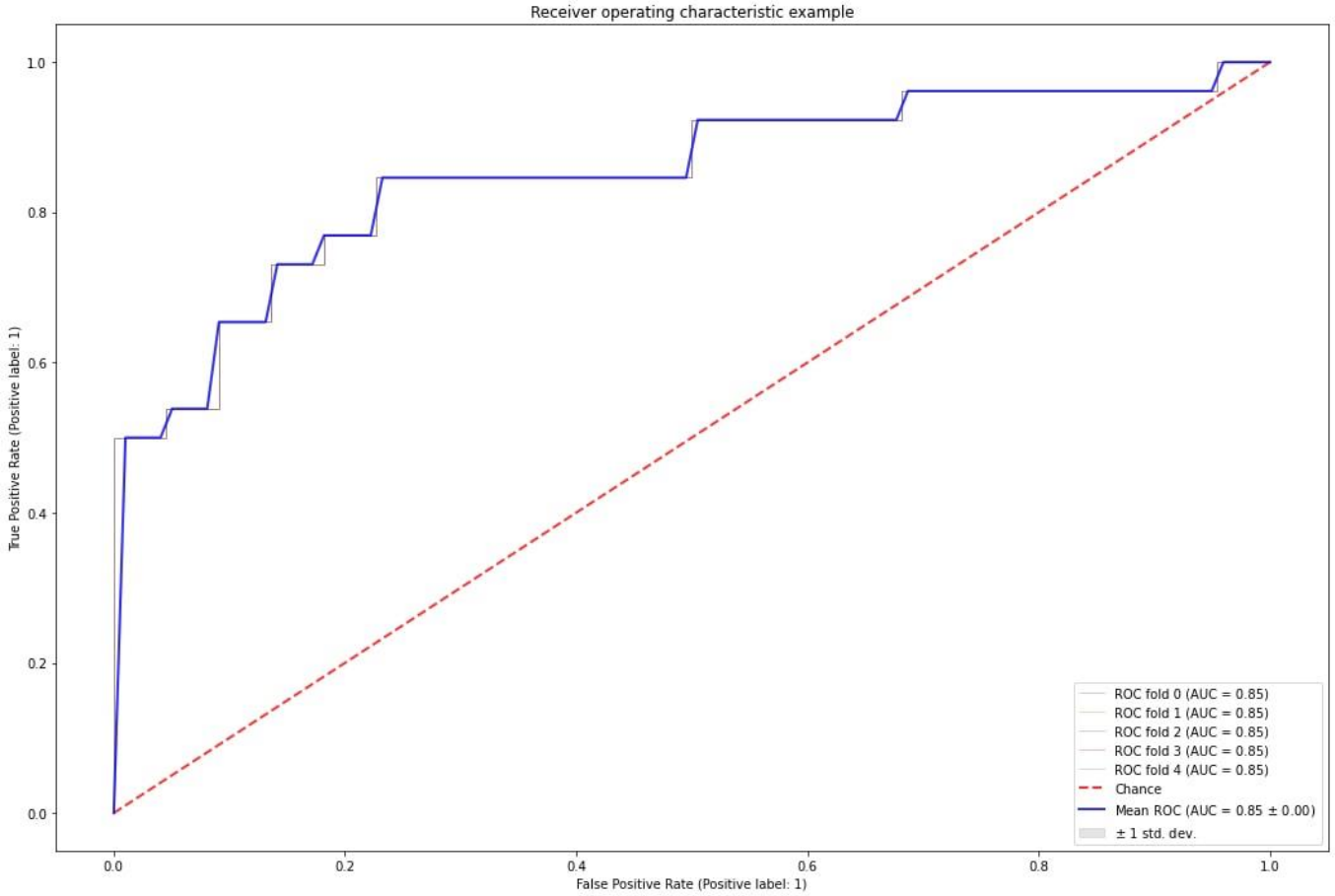
We obtain three 4x13 matrices that we reshape into a single vector of **156 features**.

We first attempt to perform **feature selection** by eliminating the least contributing features. However, the loss of information is excessive.

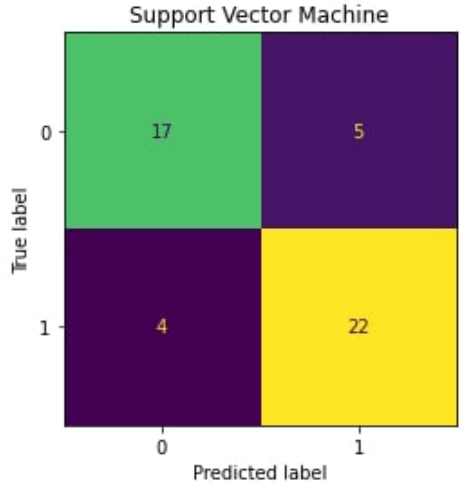
So, given the high dimensionality, we opt for a feature synthesis approach and apply **PCA** to only retain the first **2 Principal Components** as they alone explain **~89% of the total variability.**

Finally, the processed data is fed into the **different Classifiers** (SVM, Logistic Regression, Random Forest, Ensemble methods). All methods were tested with **5-fold Cross Validation** and **80-20 train/test split** stratified on the labels.

Results



After testing different Classifiers and ensembles of Classifiers, **SVM** with a **linear kernel** obtained the best results both in terms of **AUC**(85%), **Accuracy**(81%), **Precision** (81%) and **Recall**(80%).



Conclusions

To achieve better performances, we considered working on **more data**, **adding diverse tabular features** (age, sex, country of origin, presence of comorbidities or hematochemical values) and experimenting with **new image pre-processing techniques** and image-extracted features.

In conclusion, we set up a fairly **simple classification framework** that obtains **good results**, is easily **reproducible** and improved upon.

Thanks for the attention

GitHub repository for the project