**Lung Capacity Prediction in Patients with Pulmonary Fibrosis**

# ABSTRACT

Respiratory failure is the 4th leading cause of death world-wide. After 30 years of age respiratory function declines, seemingly inexorably. Exposure to cigarette smoke, infectious agents, atmospheric pollution and genetic predisposition plays important roles in accelerating this process by impinging on the ability of stem and progenitor cells to repair the respiratory epithelium.

Pulmonary fibrosis is a chronic progressive disease characterized by scarring of the lung parenchyma. As the disease progresses, the lung's ability to oxygenate blood decreases eventually leading to the death of the patient. Currently, there are no available tools to predict the prognosis of patients with pulmonary fibrosis.

The goal of this project was to develop and tune a deep learning-based algorithm that, based on patient medical information and chest CT images predicts how the lung capacity of the patient changes over time. Numerous approaches were explored, and the best performing model was an ensemble model with transfer learning from Vgg16 and made predictions with a mean absolute percent error of 4.76%.

# DATA OVERVIEW AND EDA

### Data overview

The dataset is available on Kaggle (https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression) and it was created by combining medical information data of 176 patients from several private and public hospitals. For two patients, CT images were not retrievable and thus they were discarded. For each of the remaining 174 patient the following data was available:

- **Chest CT images.** Every patient was imaged to confirm the diagnosis of pulmonary fibrosis. The number of images per patient varied from a minimum of 12 to a maximum of 1018.

- **Patient clinical information.** This included the sex, age, smoking habit (currently smoking, never smoke or ex-smoker), lung capacity measurements (FVC) and when they were taken relative to the date MRI imaging was performed. FVC stands for Forced Lung Capacity and is the amount of air expressed in milliliters a person can forcefully and quickly exhale after taking a deep breath.

### Data wrangling and features extraction

A) MRI images. Since imaging was performed in different and independent hospitals with their own machines following their own standard procedures, CT images differed from each other in:

- *Presence of a gray bounding box around the CT image.* For some patients, chest images were surrounded by a gray bounding box, which carried no meaningful information and was, therefore, removed (Exhibit 1 center panel).
- *Image resolution.* Even though 512x512 is the standard CT image resolution, some images were taken at higher resolution (e.g. 768x768) (Exhibit 1 right panel). All images were resized to the standard format of 512x512.
- *Image contrast.* The contrast of images varied from patient to patient (Exhibit 1). A well contrasted image was selected (Exhibit 1 left panel) as reference and the contrast of all the remaining images was normalized on that image.

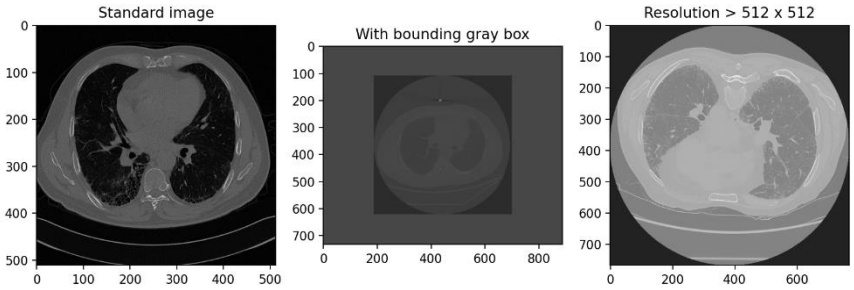


**Exhibit 1.** CT images formats

- *Imaging direction.* Some patients were imaged feet-first others head-first. The information was retrieved from the DCM files and stored in a dictionary for later use.
- *Patient arrangement on the scanning bed.* Patients can be arranged on the imaging bed in mainly 4 ways: laying on their back, chest, right side and left side. In the dataset provided, patients were imaged while laying either on their back or chest. Patient arrangement on the bed was extracted from the DCM file and stored in a dictionary for later use.
- *Number of images per patient.* The number of images per patient varied from 12 to a maximum of 1018 but some images did not show the lung of the patients but instead displayed the gut, the stomach, the neck or the head of the patient.
  The number of images *truly showing* the lung ranged for a minimum of 9 to a maximum of 227 (Exhibit 2). For each patient, we, thus, selected nine images: the closest one to the intestine, the closest one to the neck and 7 equally spaced images between them. This approach guaranteed similar representation of the lung in each patient.

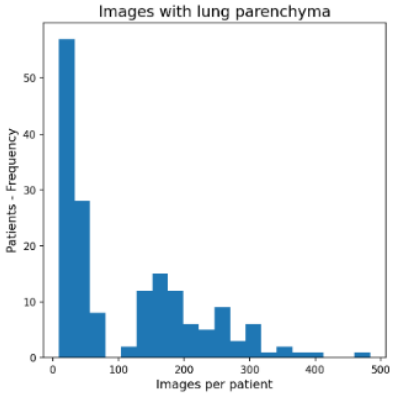


**Exhibit 2.** Images with lung per patient

The portion of the CT image showing the lung was carved out from the surrounding tissues (ribs, muscle, bone, heart and etc.) (Exhibit 3). Since pulmonary fibrosis spreads inside the lung, my thinking was that predictions made on just the lung tissue would be more accurate than predictions made on the whole image. Furthermore, removal of the black background around the carved images reduces the input size of the deep learning algorithm, potentially leading to faster training and testing.

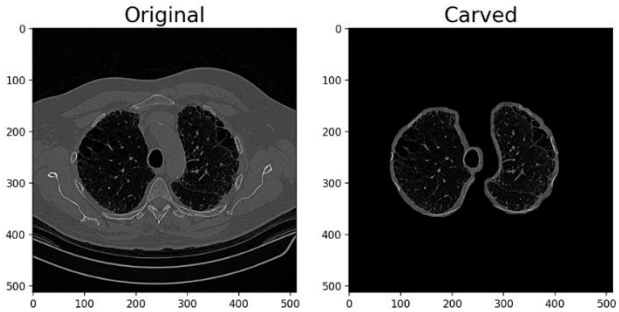


**Exhibit 3.** Original and lung carved CT images

Finally, the 9 images (carved or non-carved, with original or normalized contrast) were collaged onto 3x3 grids. Exhibit 4 shows an example of how these grids look like. The grids created with the whole images are bigger than those ones built with the carved images (1536x1536 pixels vs 1120x1460 pixels).

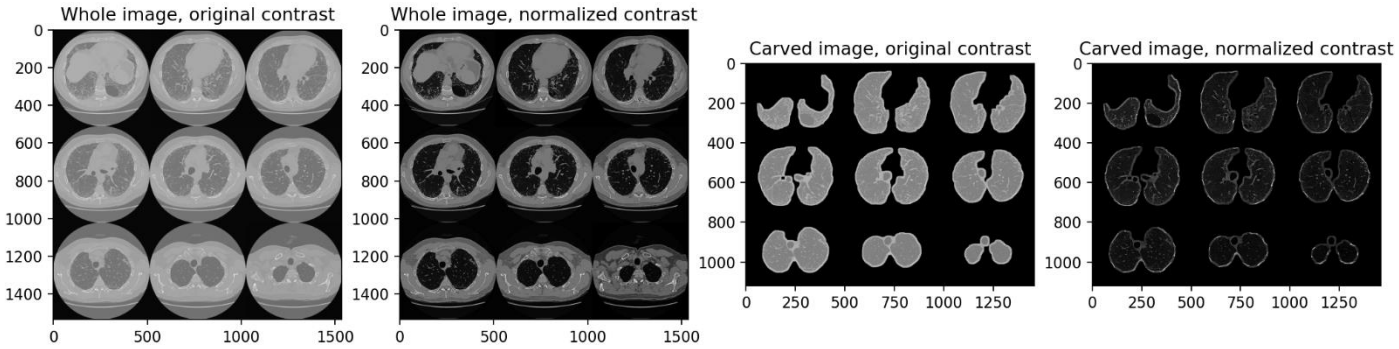


**Exhibit 4.** Example of the 3x3 images grids

B) Patient medical information. This dataset did not have missing values but had some outliers in the FVC measurements. For each patient, 6 to 10 lung capacity measurements were provided. We defined outliers those FVC values that were 2.5 standard deviation away from the mean. **Ultimately, we planned to predict the slope of the best fit line of the changes, so outliers could negatively affect the performance on new data.** Thus, outliers were removed unless they were either the first or the last measurement. We assumed that a patient's lung capacity could drastically drop on the last measurement (the patient is terminal) or after the first measurement (before the patients initiates any treatment). Exhibit 4 exemplifies these 3 scenarios.

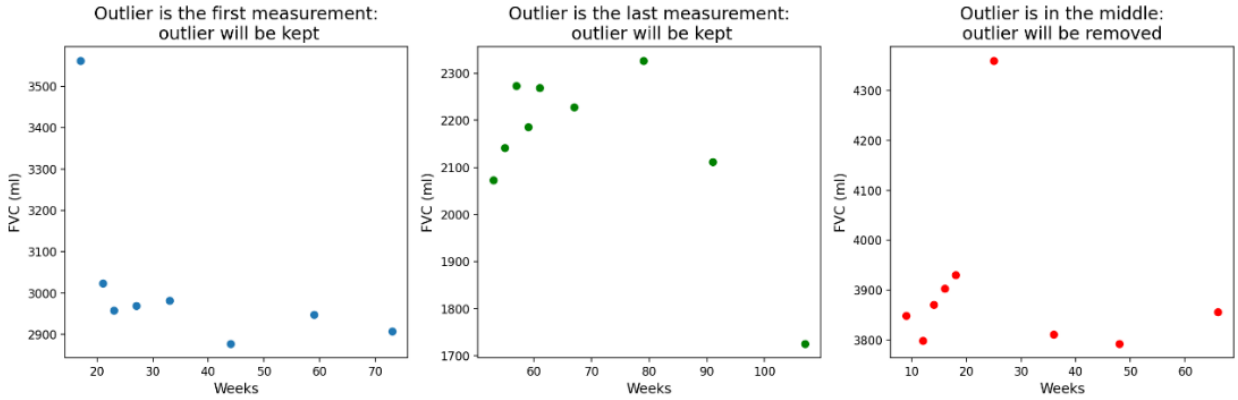


**Exhibit 4.** Outliers in the FVC measurements and how they were handled

After removal of the outliers, each patient FVC measurements were modeled with an OLS linear regressor. The slope of the best fitting line represents the dependent variable we finally want to predict.

### Exploratory data analysis

Exhibit 5 left panel shows that the age distribution of the patients is normally distributed, with a mean of 67.3 years. Most of the patients (79%) are males (Exhibit 5, center panel) and ex-smokers (67%) (Exhibit 5, right panel). This information is in line with the archetypal patient with lung fibrosis: old male with history of smoking.

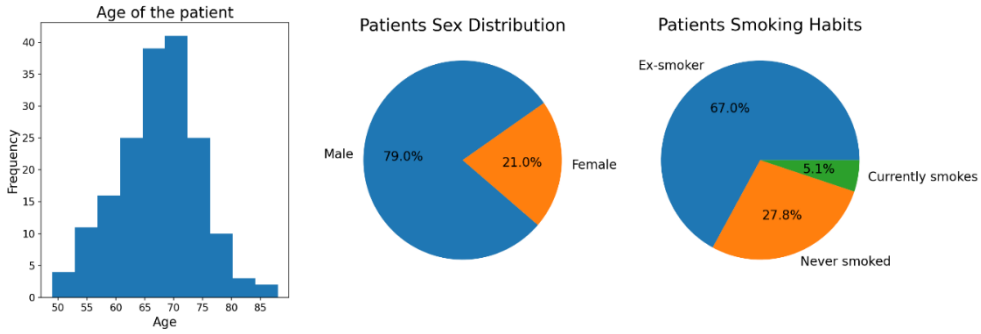


**Exhibit 5.** Patient medical information: age, sex and smoking habits

Lung capacity is typically proportional to the size of the chest and body of the patient. This fact is illustrated in Exhibit 7 where FVC measurements are plotted against the age of the patient and group by sex. Males have higher FVC values than females within the same age, as they normally have bigger bodies and chests.

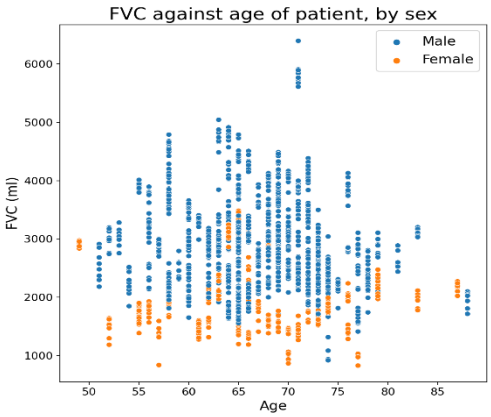


**Exhibit 7.** Patient lung capacity against age, by sex

The left panel of Exhibit 8 depicts how the patient lung capacity changed over time. For most of the patients FVC values dropped with time. This is expected since respiratory function normally starts declining at 30 years of age and all the patients in this cohort also had lung fibrosis. This is confirmed by the observation that most of the slopes of the line fitting the FVC values, is negative (Exhibit 8 right panel). However, lung capacity for a small group of patients increased over time. This could be explained by bad measurements or resolution of non-chronic respiratory diseases (e.g. cold, pneumonia, flu) which negatively affected the first few measurements.

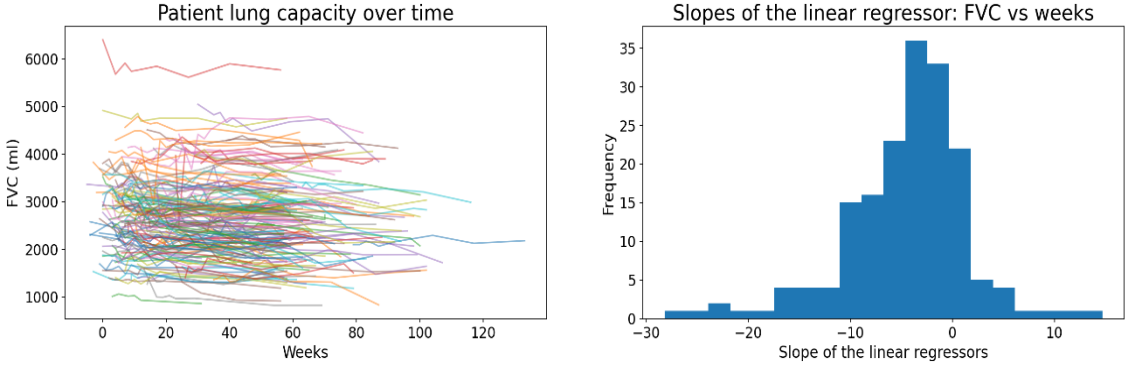


**Exhibit 8.** Patient lung capacity change over the study period

Next, using the carved images, area of the left and right lungs was calculated and plotted as a ratio against the area covered by the rest of the body (Exhibit 9). Since most of the heart is localized in the left half side of the chest, right lungs are wider and thus have higher transverse area.

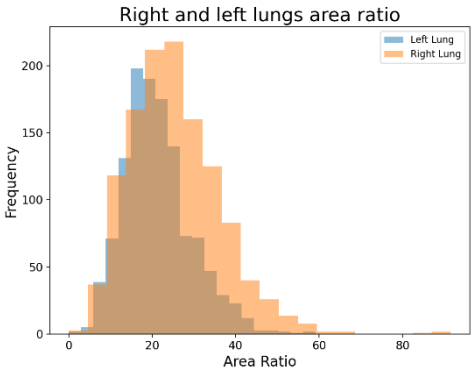


**Exhibit 9.** Area covered by right and left lung in CT images

## MACHINE LEARNING MODELING

The process to predict the patient's lung capacity consists of two steps:

1- Prediction of the slope of the line fitting the FVC values. For each patient, FVC measurements were modeled with a line. Ordinary linear square regression was deployed to estimate the slope of the best fitting line (Exhibit 10 left panel). Next, deep learning models utilizing patient's medical information, CT images or both were built and trained to predict the slope value of the line fitting the FVC measurements.

2- Calculation of the future FVC values. Once the slope was estimated, the first FVC measurement (baseline value), which is always available, as it is mandatory for diagnostic purposes, was used to calculate the intersect of the fitting line. Finally, with slope and intersect known, all the future FVC values were readily calculated (Exhibit 10 right panel).

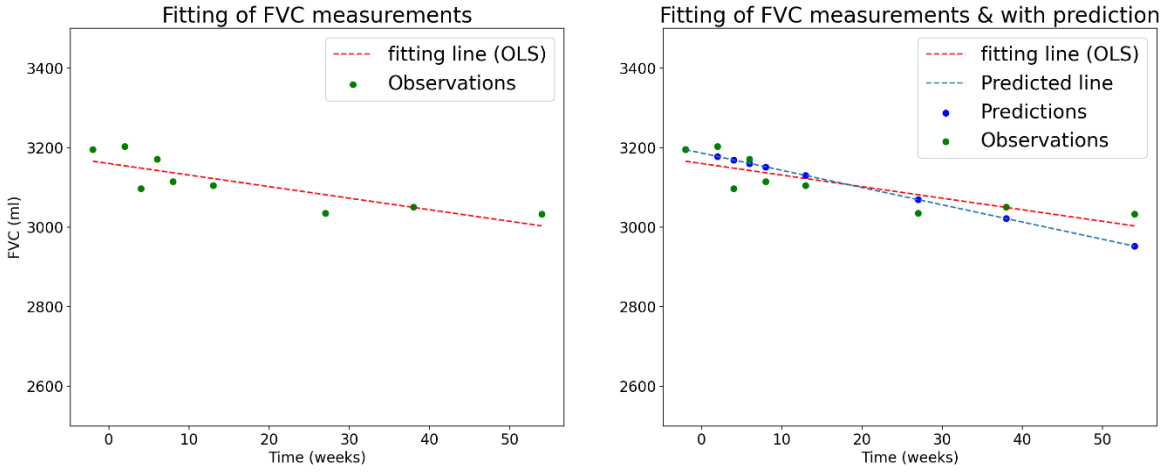


**Exhibit 10.** Patient's lung capacity prediction approach

To evaluate the coefficient of the linear model fitting the FVC measurements, several models were generated and tested:

- **Fully connected network (FCN)** with patient medical information as inputs. This model leveraged age, sex, smoking status and lungs area ratio data to predict the lung function decay rate.
- **Convolutional neural network** using patient chest CT images as input. Three separate approaches were employed:
  A) Standard approach: Images grids were fed to a convolutional neural network with two fully connected layers. In this standard approach several convolutional neural network architectures were tested and their parameters (number of filters, kernel size, etc.) tuned until the prediction error plateaued.

  B) Transfer learning. State of the art convolutional neural network models such as Vgg16 and Resnet have been trained on thousands or even millions of images in order. The first few layers of these trained models can be utilized in building new predictive models on different data. Herein, the first two layers of the trained Vgg16 were extracted and their weights frozen. The network was completed by adding 7 trainable layers (5 convolutional and 2 fully connected layers). Vgg16 is trained on color images, so the gray color CT images were converted to RGB images and downsized to 2/3 (1024x1024 pixels) to reduce computer memory usage.

  C) Autoencoders. Autoencoders have a bottleneck which forces data to be expressed in a compressed version. This approach allows the data to be represented in a compact version with little loss of information in the original data. Images grids were processed through four different autoencoders with the bottleneck "feature" layer of dimension 48x48, 96x96, 192x192 and 384x384. Exhibit 11 shows the encoding/decoding loss for each autoencoder. An example of an unprocessed image, its compressed representation and its reconstructed form is reported for each autoencoder in the *Appendix*. The 384x384 and 192x192 autoencoders produced the lowest loss and the most accurate reconstructed images among the 4 autoencoders. The compressed ('encoded') representations of CT images were used as input to a convolutional neural network whose best architecture and parameters were then identified following the approach mentioned above.

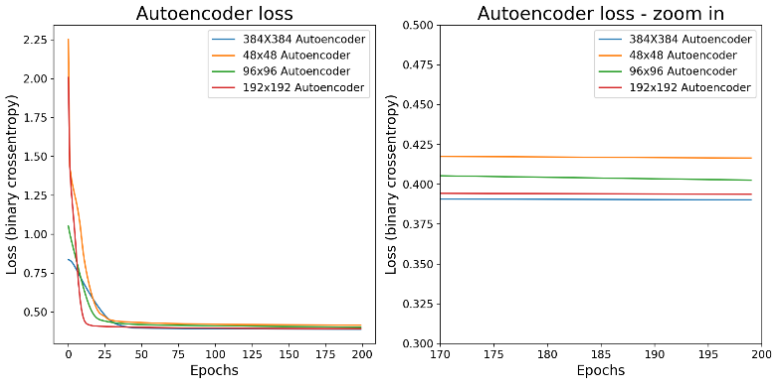


**Exhibit 11.** Loss of each of the four trained autoencoders.

- **Ensemble models.** Ensemble models can produce more accurate predictions than standalone models. We built ensemble models that made predictions based on both CT images and patient medical information. To achieve that, the prediction made the convolutional neural networks was treated as a new feature column alongside with the patient medical information. The extended patient features dataset (age, sex, smoking status, CNN slope prediction) was then fed to a fully connected neural network for the final slope prediction.

Table 1 summaries the characteristics of the models created, which data was used and their performance (mean absolute percentage error, MAPE). By comparison, when the average FVC is assigned to all the predictions, **MAPE is 28.17%**.

| Model | Input | STANDALONE MODELS<br>FVC prediction error<br>(MAPE) | ENSEMBLE MODELS<br>FVC prediction error<br>(MAPE) |
|---|---|---|---|
| **NN** | Numerical and categorical medical data | 7.19% | N/A |
| **CNN-1a** | 3x3 whole images, 1536x1536 pixels | Original contrast: 7.40%<br>Normalized contrast: 7.17% | Original contrast: 4.97%<br>Normalized contrast: 4.90% |
| **CNN-1b** | 3x3 carved images, 1120x1460 pixels | Original contrast: 7.26%<br>Normalized contrast: 7.46% | Original contrast: 4.87%<br>Normalized contrast: 5.00% |
| **CNN-2 Encoder** | 3x3 whole images, 1536x1536 pixels | Enc 48x48, Norm. contrast: 7.33%<br>Enc 96x96, Norm. contrast: 7.32%<br>Enc 192x192, Norm. contrast: 7.47%<br>Enc 384x384, Norm. contrast: 7.28% | Enc 48x48: 5.82%<br>Enc 96x96: 5.00%<br>Enc 192x192: 5.10%<br>Enc 384x384: 5.76% |
| **CNN-3 Transfer Learning** | 3x3 whole RGB images 1024x1024 pixels | Normalized contrast: 7.09% | Normalized contrast: 4.76%** |

**Table 1.** Model performance summary

The key takeaways of Table 1 are:

- Both patient medical data and chest CT images **contain key insights** for the prediction of the patient's lung capacity.
- Carving the images **does not consistently improve** the model performance. Identifying and slicing the area of the image representing the lung need a noteworthy amount of time, and computer memory, but it does not improve the prediction of the model. Thus, this step could be avoided
- Normalizing the contrast on the images **improves the model predictive power only** when the whole images are utilized. By comparing CNN-1a and CNN-1b results, it seems the best course of action is to utilize whole images and to normalize their contrast.
- Transfer learning is the **winning strategy**. The well-trained initial layers of the Vgg16 model contributed to generate the best standalone and ensemble models.

CNN-1a and CNN-1b models were not the best performing ones. Nevertheless, I created the feature map for some kernels of the convolutional layers of the model to investigate what the model "sees" in the patients' images. Exhibit 12 shows four feature maps of the last convolutional layer. **Even though it is hard to see, when closely examined, it seems that filters are detecting string-like patterns** which could be the shape of the scar tissue in the fibrotic lung.

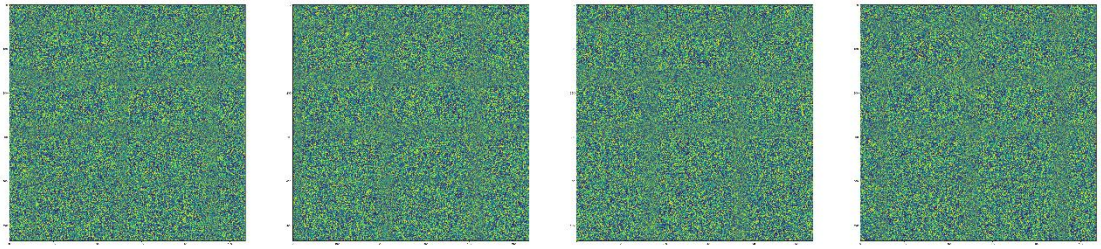

**Exhibit 12.** Feature maps of the trained CNN1a model

The results from the best model** (ensemble model with transfer learning) were further analyzed. The left panel of Exhibit 13 left panel shows the mean absolute percentage error (MAPE) for each patient. For four patients, MAPE is above 8%, with one of them approaching 12%. The error of the model seems to be inversely correlated to the mean of the FVC measurements (Exhibit 13 center panel) and, the longer the patient is enrolled in the study, the higher the error of the model (Exhibit 13 left panel). The latter observation is expected because the longer the patient is under observation the higher the chance his health status could unexpectedly change, leading to inaccurate predictions.

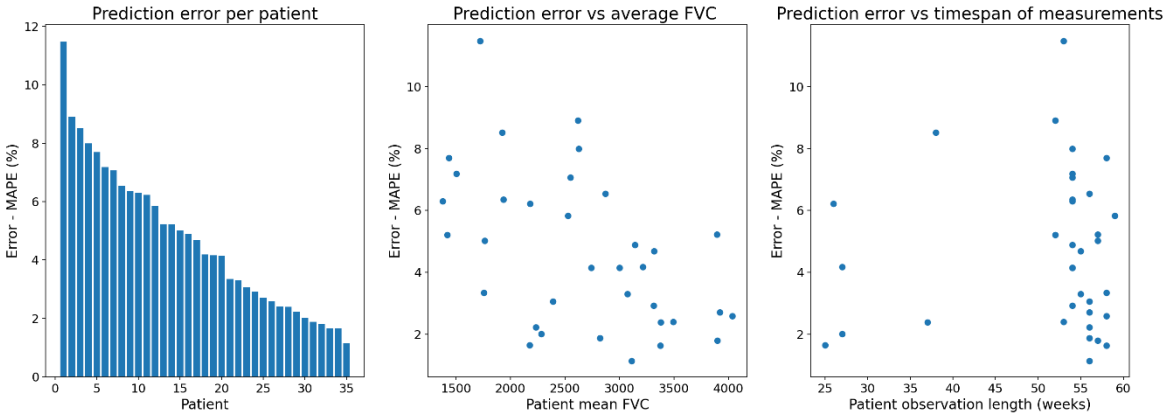


**Exhibit 13.** Error of the model analysis.

The best and worst two predictions were further analyzed (Exhibit 14). For the four plots in Exhibit 14, the y-axis was set to the same range to ease visual comparison. When the lung capacity of the patient drops linearly overtime, the model performs well. However, it fails to predict positive and negative erratic changes of the lung function.

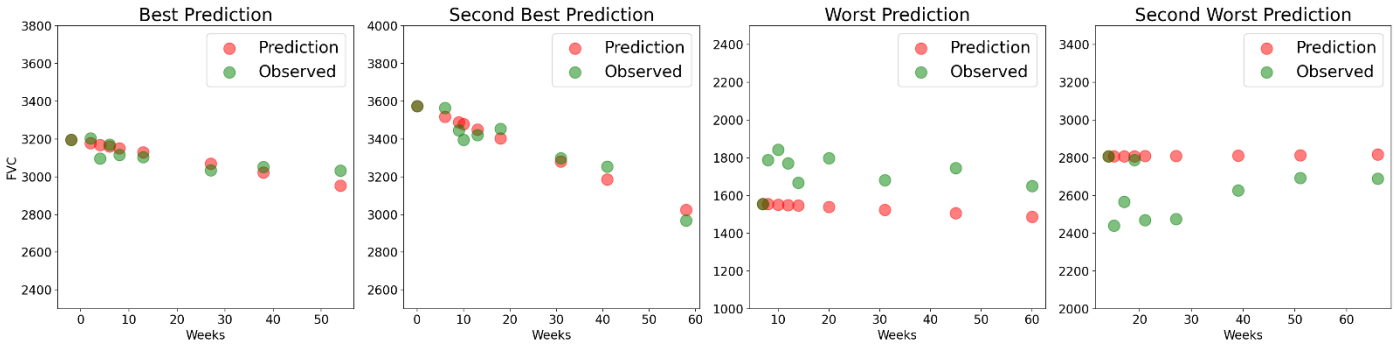


**Exhibit 14.** Best and worst predictions

## Conclusion and future directions.

Both patient medical information and chest CT images contain key insights for the prediction of the patient's lung capacity. Of all the tested approaches, the ensemble model transfer learning from the pretrained Vgg16 model is the winning strategy. Deeper analyses showed that the model performs well when lung capacity decays almost linearly but underperforms otherwise. In the clinical setting, one way to correct for this issue,

is to utilize the last FVC measurement (or the average of last n measurements) as baseline value for the following predictions. This approach will allow the model to better adapt to unexpected drastic changes of the patient lung capacity.

The model could be improved by including more information about the patient. For instance, there is no information on the pharmacological treatment (and when it was initiated), presence of comorbidities (such as diabetes, cardiovascular disease, other chronic diseases) and other general information (blood pressure, weight, body mass index). Furthermore, more images per patient could help improving model performance. If the number of images is high enough, the whole 3D volume of the lung could be processed producing the most comprehensive solution. Finally, the size of the dataset could be bigger. 176 is objectively very small number, given the complexity of the problem we are trying to solve.

In conclusion, with a mean absolute percentage error ranging from 4.76% the best model produces satisfying results. However, since the worst predictions are obtained for patients with drastically drops in lung function, the implementation of this model should be carefully well-thought-out and should not substitute the assessment of the patient's physician.

## Appendix

A) Encoded representation and reconstructed image for the 4 autoencoders