PROGRESS PREDICTION 4°/5° WEEK

Code Review

Extract Features

Generate csv file with informaton generated from images

1. Metadata extracted SliceThickness and PixelSpacing

2. Make Lung mask
   a. Normalize image → remove mean and divide by std
   b. Renormalize washed images → sub light/dark pixels with mean
   c. K-means to separate foreground and background
   d. Erosion → eliminate noise/small details with a 3x3 filter
   e. Dilation → reconstruct principal areas through a 8x8 filter
   ....

Code Review

Extract Features

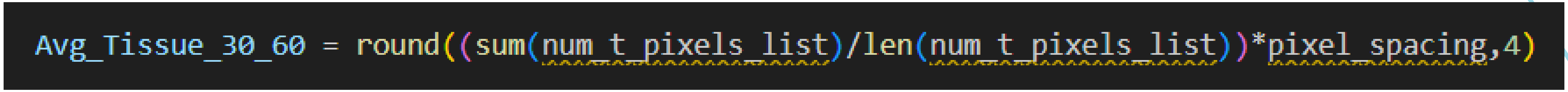
....

a. Label creation (skimage) → assign labels for each portion
b. Compute geometrical attributes (area,bounding box)
c. Select good bounding boxes → eliminate too big/small areas
d. Fill lung masks → 1 for lungs, 0 elsewhere
e. Compute lung area
f. Calculate tissue mask and extract features (lung without border)

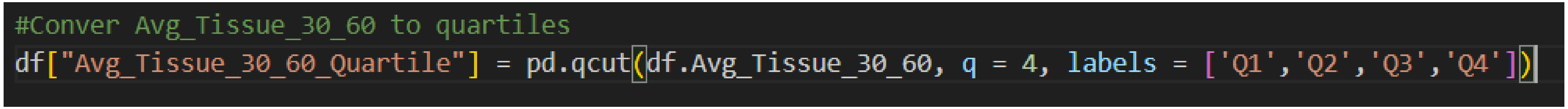
3. Join extracted features to metadata and known data

Code Review

Quantile definition



```
Avg_Tissue_30_60 = round((sum(num_t_pixels_list)/len(num_t_pixels_list))*pixel_spacing,4)
```



```
#Conver Avg_Tissue_30_60 to quartiles
df["Avg_Tissue_30_60_Quartile"] = pd.qcut(df.Avg_Tissue_30_60, q = 4, labels = ['Q1','Q2','Q3','Q4'])
```

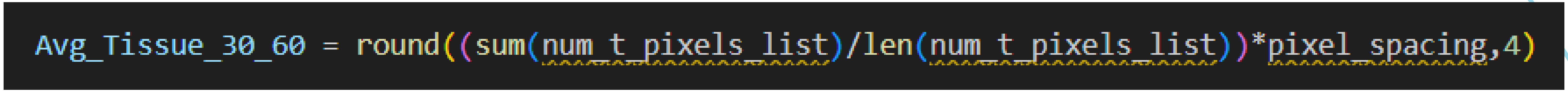
Uses Avg_tissue_30_60 to define quantile groups and define categorical values.
Computed through:

- **num_t_pixels_list : list of the number of tissue pixels detected in image slices between 30% and 60% of the lung height**
- **pixel_spacing : metadata**

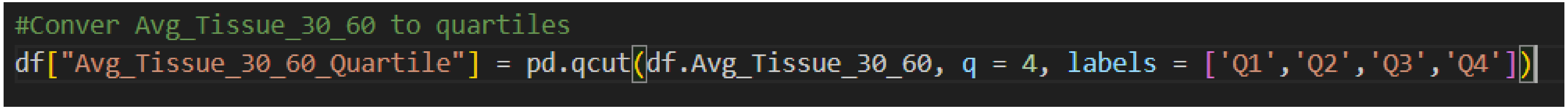
So it’s the average tissue area (in mm^2).

Code Review

Quantile definition



```
Avg_Tissue_30_60 = round((sum(num_t_pixels_list)/len(num_t_pixels_list))*pixel_spacing,4)
```



```
#Conver Avg_Tissue_30_60 to quartiles
df["Avg_Tissue_30_60_Quartile"] = pd.qcut(df.Avg_Tissue_30_60, q = 4, labels = ['Q1','Q2','Q3','Q4'])
```

“pd.qcut” used to divide data into 4 groups with the same amount of data 4 groups based on percentiles (25,50,75).

labels=['Q1','Q2','Q3','Q4'] assigns quartile names:

- **Q1: lowest 25% of average tissue areas**
- **Q2: 25–50%**
- **Q3: 50–75%**
- **Q4: top 25% (largest average tissue areas)**

Code Review

Modeling 1

For each patient p in the train set:

- Fits a linear regression and saves the slope (a), tab values and patient

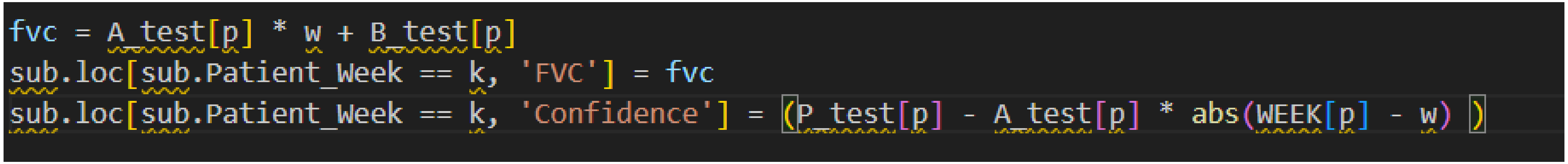
Generates 5 folds and split patients between these 5 folds.
For each iteration chooses 4 for training and 1 for validation.

Per iteration it builds a new efficient model (so for each iteration it trains the model on a slightly different training set).
Per iteration each patient in the test set, gets slices and tabular values and predict a slope for each slice , choosing the slope through the quantile selected for that fold.

Code Review

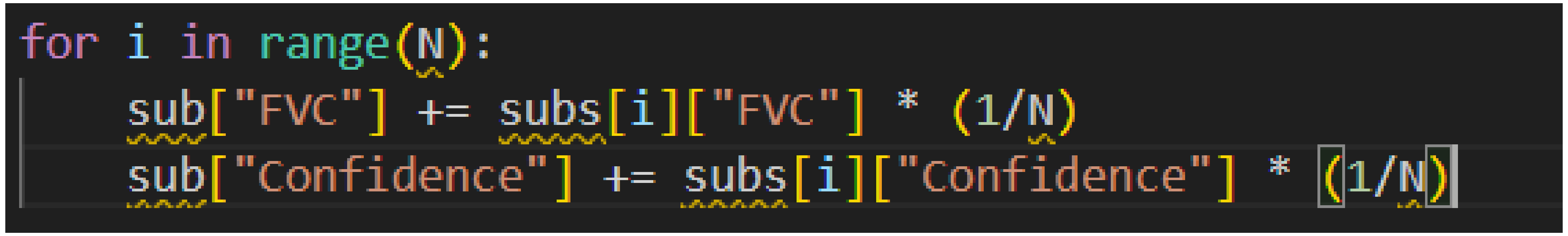
Modeling 1

Having the predicted slope, we can predict the FVC and Confidence for the week defined in the sample_submission csv.
How:



```
fvc = A_test[p] * w + B_test[p]
sub.loc[sub.Patient_Week == k, 'FVC'] = fvc
sub.loc[sub.Patient_Week == k, 'Confidence'] = (P_test[p] - A_test[p] * abs(WEEK[p] - w) )
```

In the end we will have a different prediction for each iteration and an average will be made.



```
for i in range(N):
    sub["FVC"] += subs[i]["FVC"] * (1/N)
    sub["Confidence"] += subs[i]["Confidence"] * (1/N)
```

Code Review

Preparation data

Prepare data:

- Add a train/test/val column
- Add minimum week column (earliest visit for patient)
- Baseline FVC column
- Baseline Percent column
- Add column to indicate time passed from baseline visit
- One-hot encoder for Sex and SmokingStatus
- Add image features extracted from image

Merge all data, handle outliers and noise and normalize.

Code Review

Modeling 2

The models final output is formed by three values:

[y_lower, y_pred, y_upper]

Representing the lower quantile,median and upper quantile estimates of FVC for a patient at a given week.

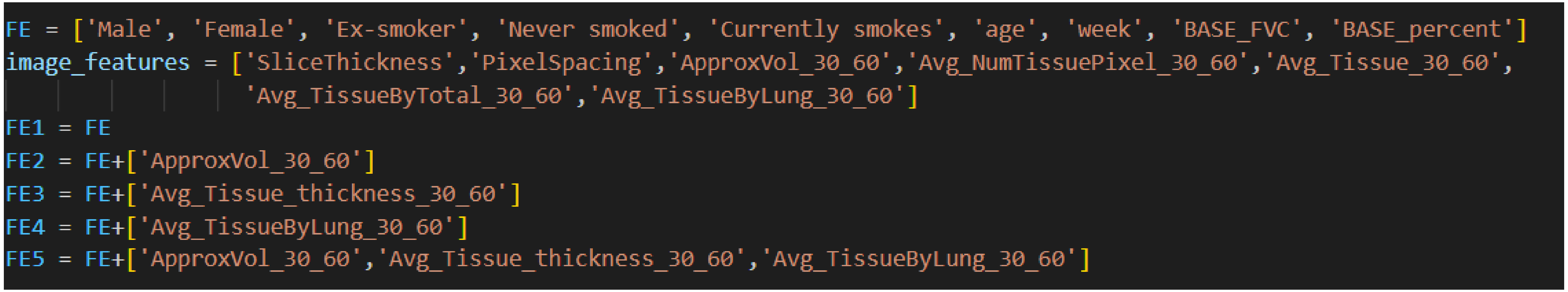
The model uses a combined loss (mloss):

- **qloss → encourages predictions for each quantile to bracket the true value correctly**
- **score → approximates the laplace log-likelihood**

Code Review

Modeling 2

5 Neural Networks, each work on a slightly different feature set:



```
FE = ['Male', 'Female', 'Ex-smoker', 'Never smoked', 'Currently smokes', 'age', 'week', 'BASE_FVC', 'BASE_percent']
image_features = ['SliceThickness','PixelSpacing','ApproxVol_30_60','Avg_NumTissuePixel_30_60','Avg_Tissue_30_60',
                  'Avg_TissueByTotal_30_60','Avg_TissueByLung_30_60']
FE1 = FE
FE2 = FE+['ApproxVol_30_60']
FE3 = FE+['Avg_Tissue_thickness_30_60']
FE4 = FE+['Avg_TissueByLung_30_60']
FE5 = FE+['ApproxVol_30_60','Avg_Tissue_thickness_30_60','Avg_TissueByLung_30_60']
```

Collects all predictions and search for optimal ensemble weights - in a brute-force way - across the 5 models.

Code Review

Modeling 2

The final ouput FVC and Confidence is given by:

- FVC : median value (y_true)

- Confidence: y_upper - y_lower

Then finally it blends the predictions to the first model:

- 40% image-based model
- 60% metadata model

Comparison to other approaches

- 5th Place:

Small network with only tabular data
Inputs → [WeekInit, WeekTarget, WeekDiff, FVC, Percent, Age, Sex, CurrentlySmokes, Ex-smoker, Never Smoked]

- 6th Place:

Each measurement in the dataset is treated as if it were a baseline measurement. A new feature week_passed is created and extracted image features as base data. Used 5 models and weighted them
[Lasso, Ridge, ElasticNet, SVM, NN] = [0.68573749, 0., 0., 0.07551167, 0.23750526]

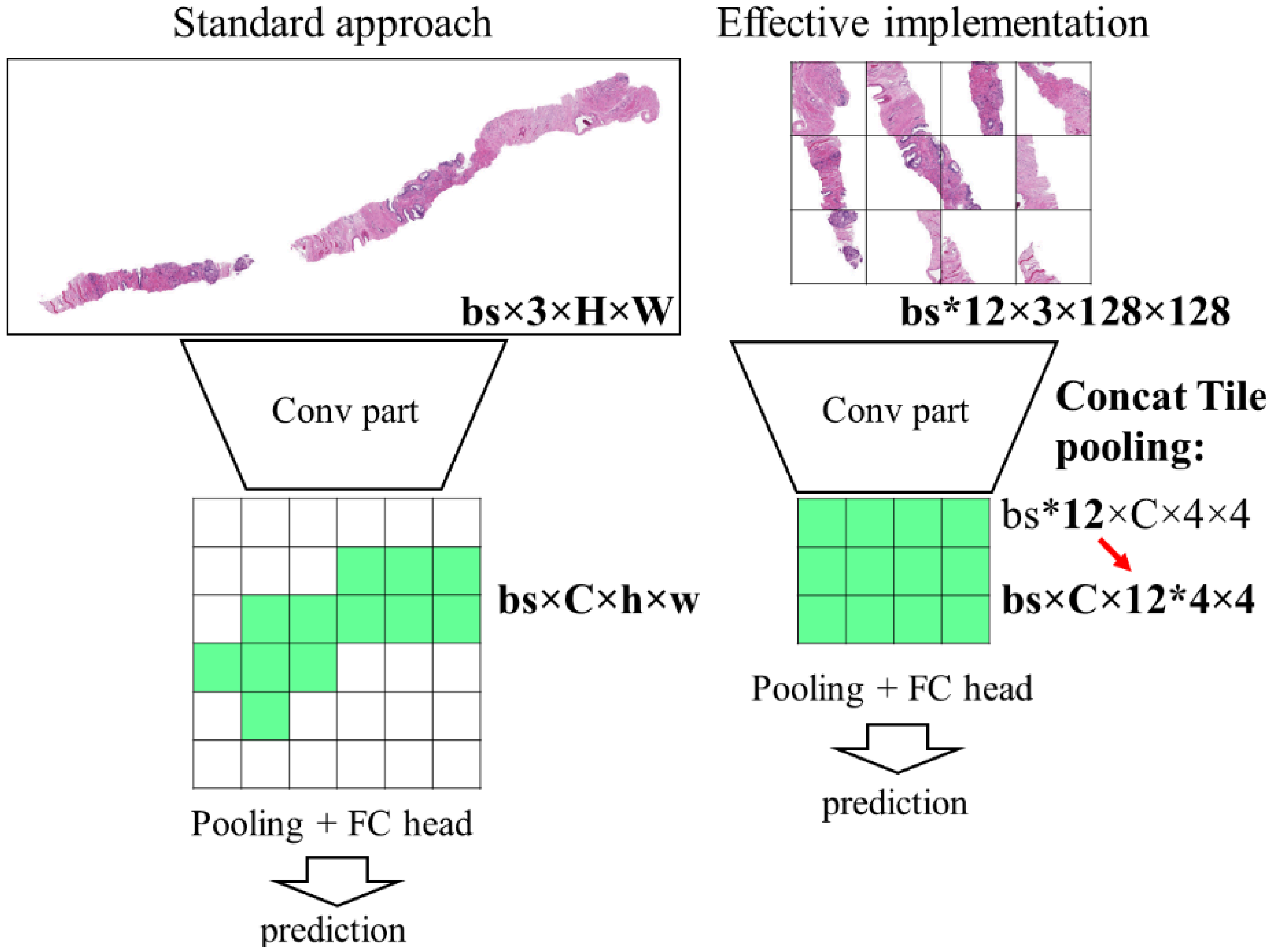
- 9th Place:

Use of Concatenate Tile Pooling approach for 2D CT scans, aggregates information across multiple CT layers and assigns a single label to the entire scan.

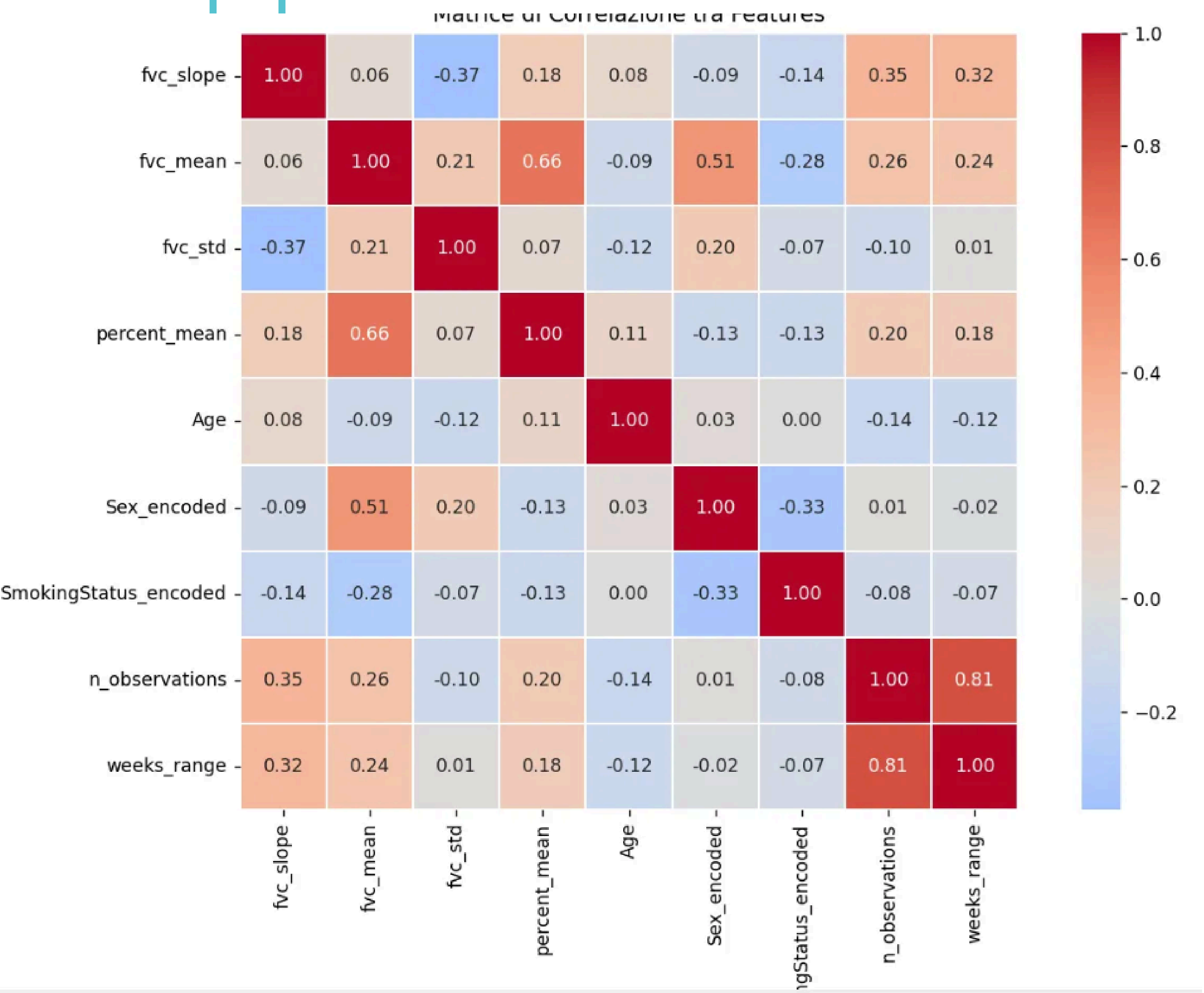
Concatenate Tile Pooling

Instead of passing an entire image as an input, N tiles are selected from each image based on the number of tissue pixels and passed independently through the convolutional part.

The outputs of the convolutional part is concatenated in a large single map for each image preceding pooling and FC head .



PREDICTION FOR WEEK 0



Matrice di Correlazione tra Features

	fvc_slope	fvc_mean	fvc_std	percent_mean	Age	Sex_encoded	SmokingStatus_encoded	n_observations	weeks_range
fvc_slope	1.00	0.06	-0.37	0.18	0.08	-0.09	-0.14	0.35	0.32
fvc_mean	0.06	1.00	0.21	0.66	-0.09	0.51	-0.28	0.26	0.24
fvc_std	-0.37	0.21	1.00	0.07	-0.12	0.20	-0.07	-0.10	0.01
percent_mean	0.18	0.66	0.07	1.00	0.11	-0.13	-0.13	0.20	0.18
Age	0.08	-0.09	-0.12	0.11	1.00	0.03	0.00	-0.14	-0.12
Sex_encoded	-0.09	0.51	0.20	-0.13	0.03	1.00	-0.33	0.01	-0.02
SmokingStatus_encoded	-0.14	-0.28	-0.07	-0.13	0.00	-0.33	1.00	-0.08	-0.07
n_observations	0.35	0.26	-0.10	0.20	-0.14	0.01	-0.08	1.00	0.81
weeks_range	0.32	0.24	0.01	0.18	-0.12	-0.02	-0.07	0.81	1.00

Clustering

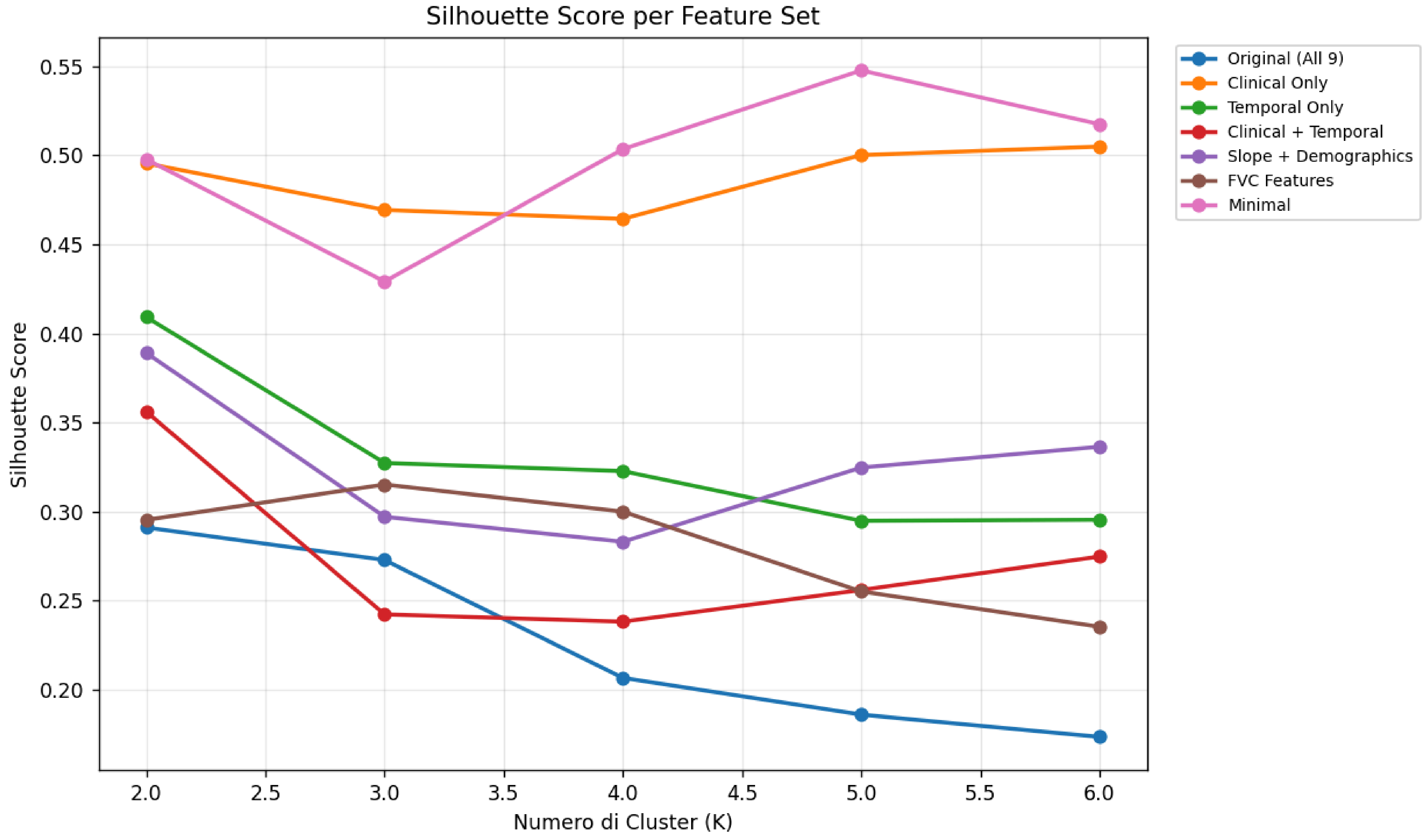
Divide patients into clusters and adapt each linear regression to the cluster to define similar characteristics and get informations from multiple individuals

PREDICTION FOR WEEK 0

Different subsets of features starting from 9 features :

- fvc_slope
- fvc_mean
- fvc_std
- percent_mean
- Age
- Sex_encoded
- SmokingStatus_encoded
- n_observations
- weeks_range

Best one: Minimal
(fvc_slope, sex, smokingstatus)



PREDICTION FOR WEEK 0

Formed clusters divide patients in:

Cluster	N_Patients	FVC_Slope_Avg	FVC_Mean	Age_Avg	Male_%	Smoking_Mode	N_with_Week0	FVC_intercept_mean
0	87	-1.94	2903.85	67.79	100	Ex-smoker	5	2963.51
1	23	-4.18	1765.79	66.61	0	Never smoked	1	1923.01
2	29	-14.12	2808.48	66.31	100	Ex-smoker	4	3225.92
3	23	-3.86	2936.48	67.09	100	Never smoked	6	3068.14
4	14	-2.37	2042.42	67.29	0	Ex-smoker	2	2122.44

Cluster 0: FVC slope → -1.94

- Only Male
- 80 ex-smoker
- 7 currently smokes
- FVC Decline slow/stable
- MAE 1174.14
- R^2 0.004

Cluster 1: FVC slope → -4.18

- Only Female
- 23 Never smoked
- FVC Decline moderate
- MAE 167.79
- R^2 0.003

Cluster 2: FVC slope → -14.12

- Only Male
- 26 ex-smoker
- 3 never smoked
- Rapid FVC decline
- MAE 456.58
- R^2 0.055

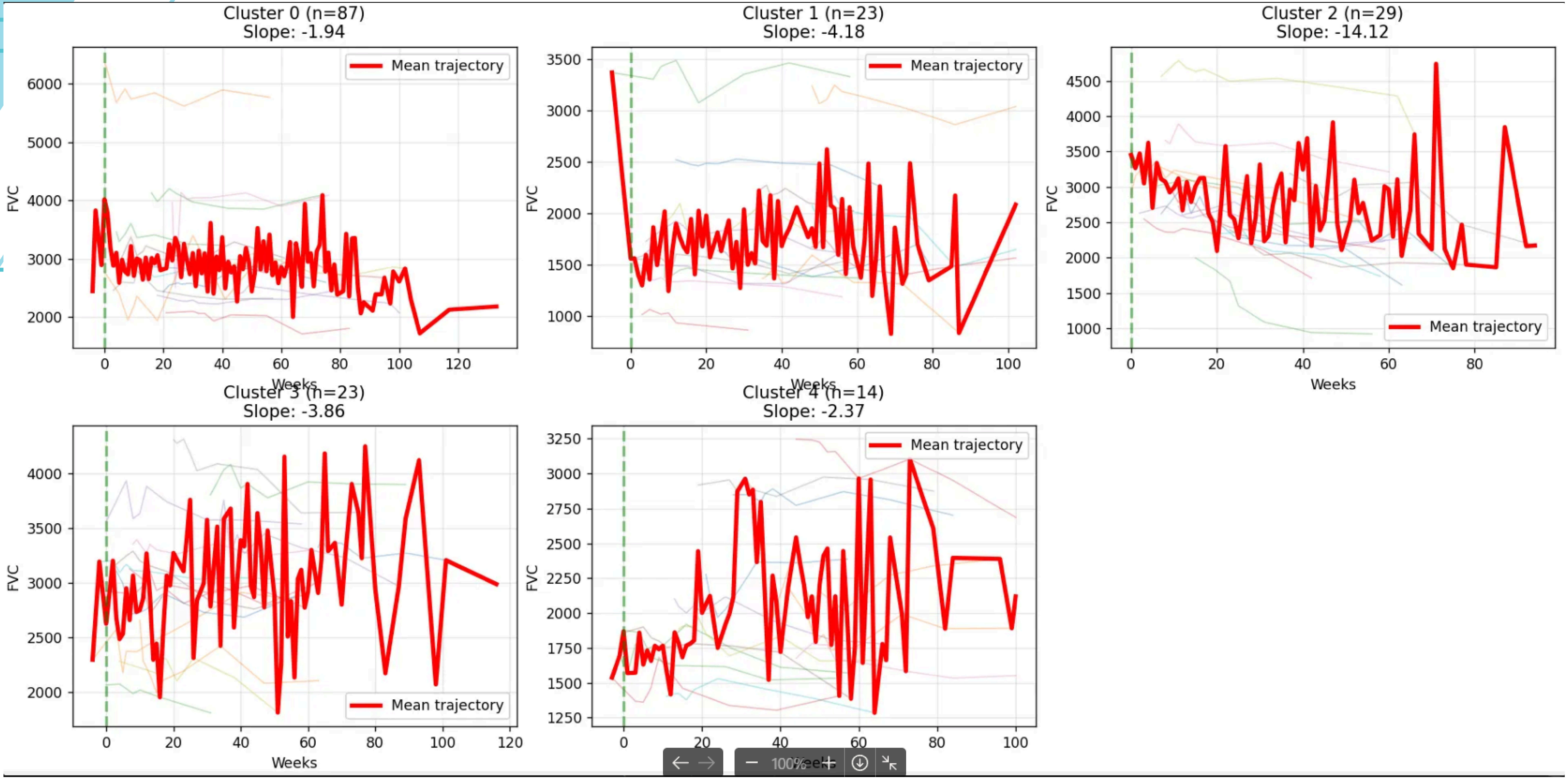
Cluster 3: FVC slope → -3.86

- Only Male
- 23 never smoked
- FVC Decline moderate
- MAE 532.50
- R^2 0.054

Cluster 4: FVC slope → -2.37

- Only Female
- 12 ex-smoker
- 2 currently smokes
- FVC Decline slow/stable
- MAE 68.46
- R^2 0.087

Validation made on the 18 patients with week 0 so the values of MAE and R^2 are very apporoximate.


Cluster 0 (n=87)
Slope: -1.94
Mean trajectory
FVC
6000
5000
4000
3000
2000
0
20
40
60
80
100
120
Weeks
Cluster 1 (n=23)
Slope: -4.18
Mean trajectory
FVC
3500
3000
2500
2000
1500
1000
0
20
40
60
80
100
Weeks
Cluster 2 (n=29)
Slope: -14.12
Mean trajectory
FVC
4500
4000
3500
3000
2500
2000
1500
1000
0
20
40
60
80
Weeks
Cluster 3 (n=23)
Slope: -3.86
Mean trajectory
FVC
4000
3500
3000
2500
2000
0
20
40
60
80
100
120
Weeks
Cluster 4 (n=14)
Slope: -2.37
Mean trajectory
FVC
3250
3000
2750
2500
2250
2000
1750
1500
1250
0
20
40
60
80
100
Weeks
100%

Comparison between methods:

- Cluster based on FVC_slope, Sex, Smoking Status
  - MAE : 622.04

- Individual (based on individual slope → not cluster)
  - MAE : 127.62

- Cluster based on all 9 features
  - MAE : 536.12

The MAE increases for the first method respect to where we have 9 features in the clusters, but the clusters become more interpretable.

Best approach between the three methods

- Hybrid Approach
  - Use individual method for patients with a good fit,
  - Use optimized clusters for patients with variable data , and far from week 0

Use individual method if MAE% value better than cluster:

- MAE%< 5 and distance from week 0 < 15 → personal high confidence
- MAE%<10 → personal medium confidence
- MAE%<12 → personal low confidence

else use cluster estimate → cluster optimized

With this new method:

- **Personal high confidence (n=87)**
  - **Avg R^2 : 0.450**
  - **Distance week0 : 6.7 weeks**
  - **Avg MAE: 67**
- **Personal medium confidence (n=49)**
  - **Avg R^2 : 0.358**
  - **Distance week0 : 21.1 weeks**
  - **Avg MAE: 98.0**
- **Cluster optimized (n=20)**
  - **Avg R^2 : 0.479**
  - **Distance week0 : 46.5 weeks**
  - **Avg MAE: 74.2**
- **Personal low confidence (n=1)**
  - **Avg R^2 : 0.394**
  - **Distance week0 : 13.0 weeks**
  - **Avg MAE: 128.3**
  -

Problem → Is it Reliable?

We work with predicted values of Week 0.
We'll have a final prediction of the progression based on a starting prediction, there could be a big error propagation.

Possible other ways:

- Work with predicted values at week 0 but weight data, giving much more importance to the 18 patients we know and less to the ones predicted (especially the ones with low confidence)

- Eliminate FVC at week 0 for everyone, use only CT scans as baseline (extracting from there the FVC?) therefore final output cannot be decline based on baseline FVC but a definite value

- Incorporate the prediction of the FVC baseline at time as an output of the model itself, so final output (FVC baseline, FVC 1year, FVC 2year, Progression Yes/No)