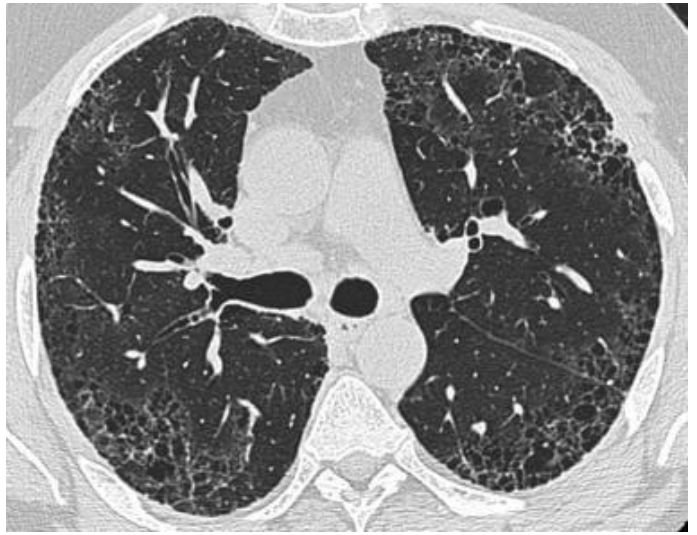
What is Pulmonary Fibrosis?

- Pulmonary fibrosis is a lung disease that occurs when **lung tissue becomes damaged and scarred**
- Patients diagnosed with pulmonary fibrosis experience symptoms like shortness of breath and as the condition progresses, **overall lung function of the patient declines**.



Problem and Motivation

Problem:

- wide variation in how the disease progresses in each patient
- Unpredictability in rate of decline – Leads to anxiety for patients and causes problems in treatment planning

Need to Solve – Why predict lung function decline ?:

- It would significantly aid patients and doctors to make much more informed medical decisions and build better treatment strategies
- Understanding prognosis would help in managing the disease better

Literature Survey:

1. Sampurna Mandal et al. has used three models, they have shown the comparisons among various ML (Machine learning) models' performance to analyse disease Progression. Among Quantile Regression, Ridge Regression and ElasticNet reported that the ElasticNet model has given the best results

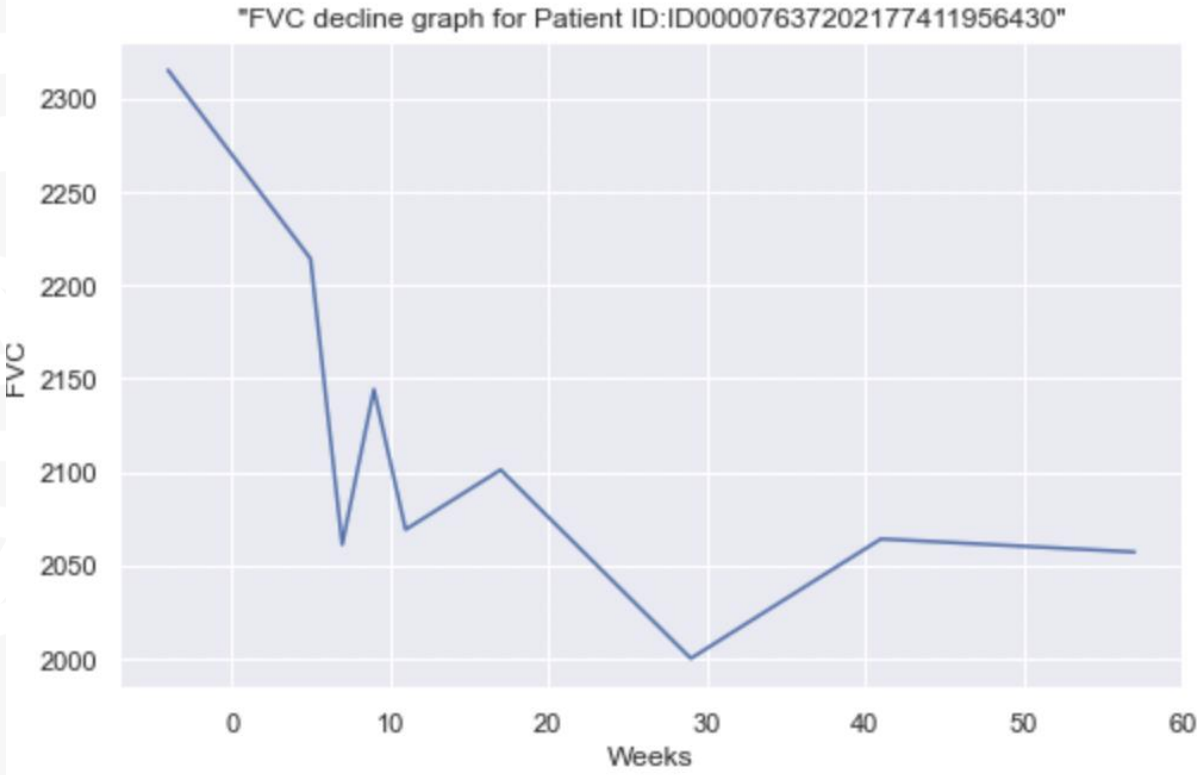
2. Alexander Wong et al. proposes a new neural network named Fibrosis-Net to predict FVC, they claim that their model has explainability driven performance as it was able to exhibit correct decision-making behavior by leveraging clinically-relevant visual indicators in CT images when making predictions on pulmonary fibrosis progress

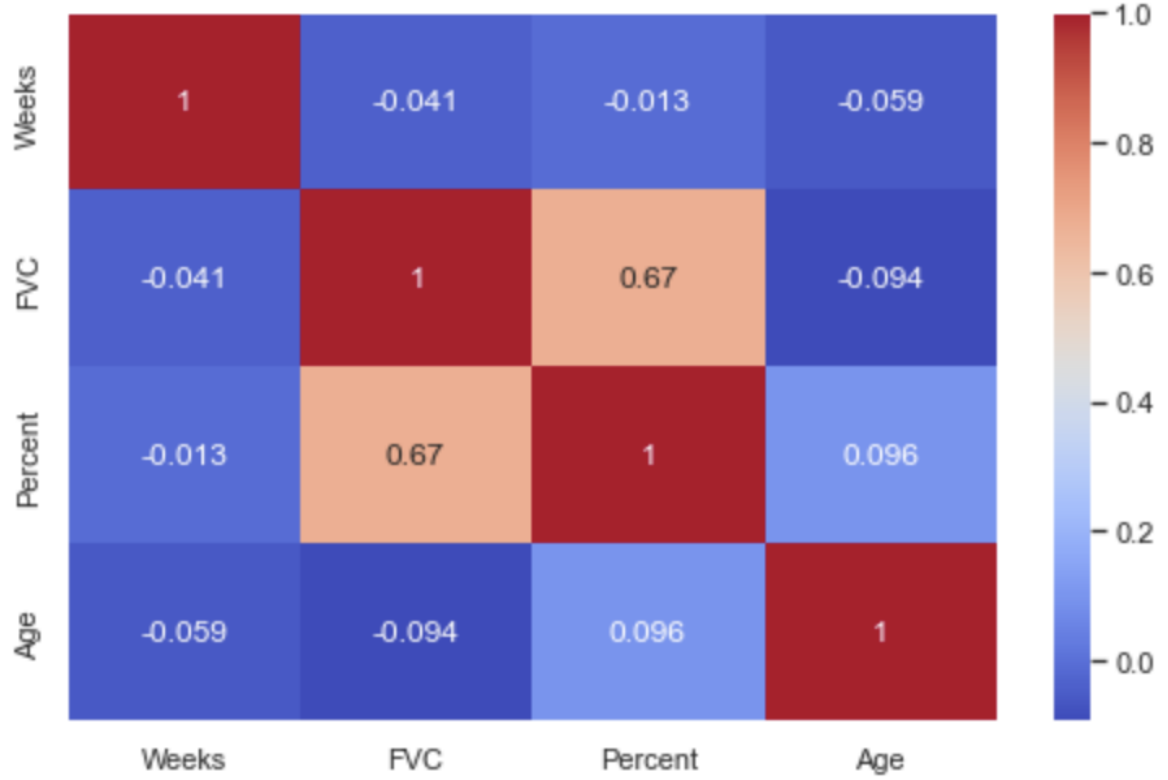
3. Anju Yadav et al. uses the honey combing pattern in the lung images to study the lungs and used their own Neural network called FVC-Net to predict the FVC, they applied image processing (edge detection) techniques to calculate the degree of honey combing and edge detection which can tell about the lung deterioration and subsequently to predict FVC

Data Description:

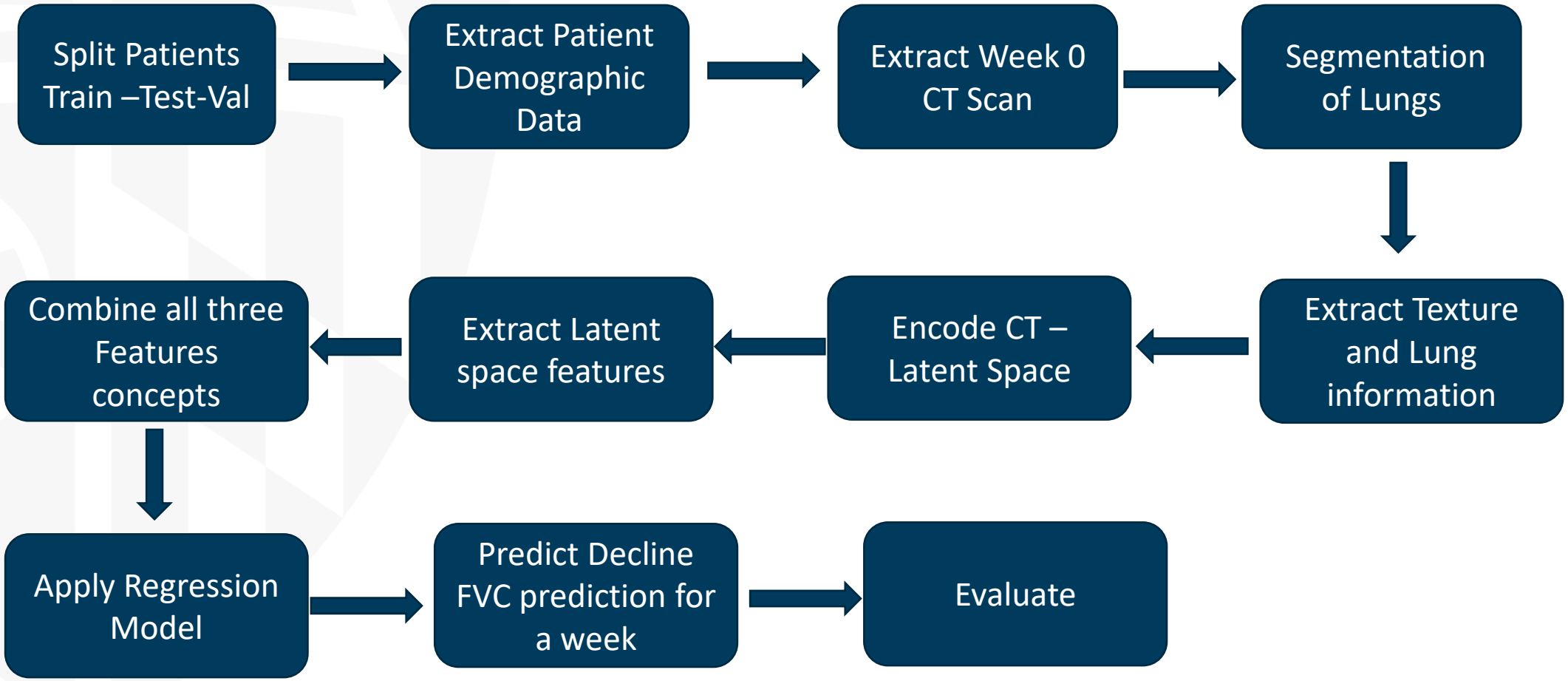
- The data provided is collected in a **longitudinal study**
- **Two kinds of Data were provided:**
- **Structured data:**
  - Excel sheet files: Demographic Information of the patients (Patient ID, week, FVC(Forced Vital Capacity), Age, Sex, Percent, and Smoking status)
- **Unstructured Image Data:** CT scan of the Patients
  - The CT scan is taken once at the beginning of the study and FVC measurements are taken on ongoing weeks (weeks need not be consecutive also) through out the study
  - As CT scan is 3d image, there were different number of slices for each patient as there could be devices with different resolution, hence we sampled 30 slices for each.(For scans <30 slices we retained them as they are)

Data Exploration and Visualization:





Approach - Overall Pipeline to predict decline



Approach – Three Stages

- Extract Demographic features
- Extract handcrafted features
- Extract latent features

Stage 1 – Demographic features

The demographic features for each patient are provided in excel format. For Each Patient ID we get the following features

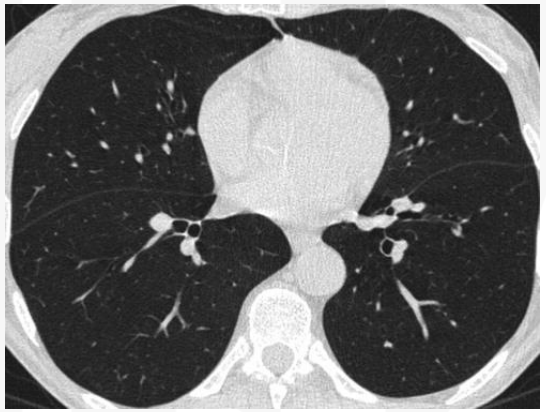
- Age
- Gender
- Smoking Status
- Percentage Health

Gender and Smoking status being categorical – We encode them using dummy variables

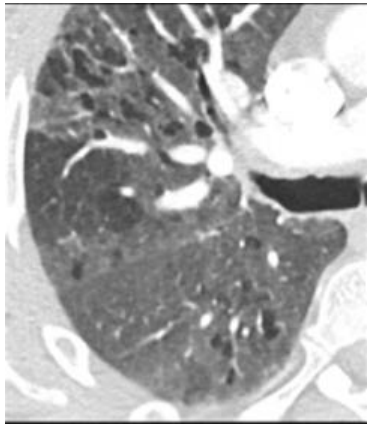
Stage 2 – Generating Handcrafted Features

What do Radiologists /Doctors Look into CT for diagnosing Pulmonary Fibrosis:

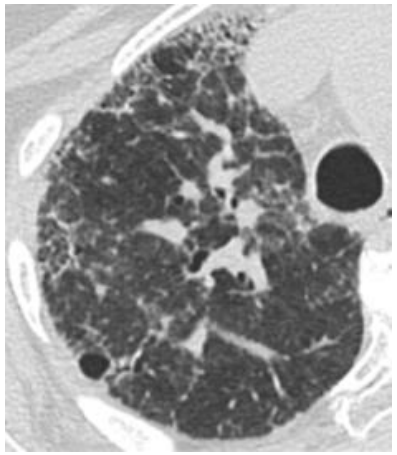
- Glassy visuals
- Honeycombing structures



Normal Lung – CT mid slice



Glassy Lung – CT mid slice



Honey comb like structures

What are we trying to extract ?

- We see clear Textural variations – Can we extract texture?

Stage 2 – GLCM Features and Lung Anatomy Information

Gray-Level Co-Occurrence Matrix (GLCM) – Statistical method to examine spatial relationships of pixels

Generates :

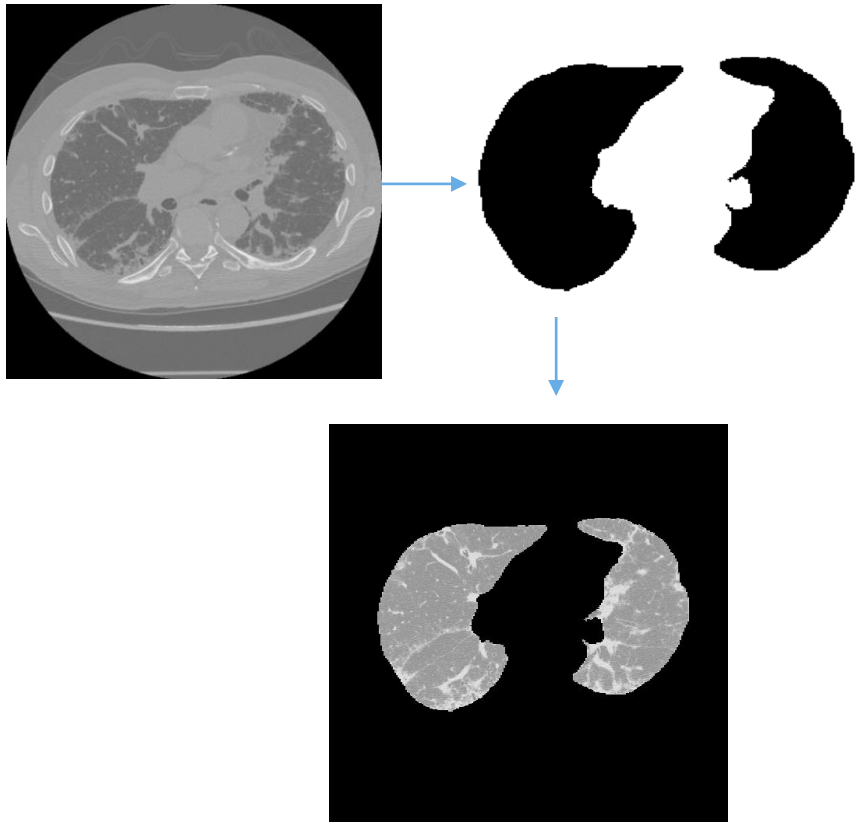
- Contrast , Correlation , Entropy and Momentum features

Stage 2 Overall Process:

- Get CT Scan – Scan has n slices
- Segment lungs in each slice – Using pretrained UNET architecture already available *
- Apply GLCM on each extracted lungs
- Divide scans into top – Middle – Lower Lobes of lungs – Get average GLCM for top, middle and lower lobes
- Extract lung volume – Useful since we are predicting FVC

*- Git - JoHof/lungmask

Segmentation result Sample



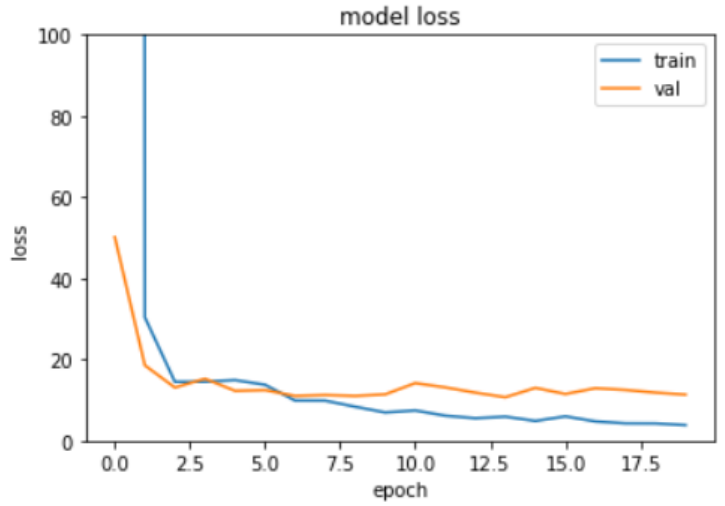
Stage 3 – Encoding and Latent Space Features

- Unet encoder part – inspired architecture
- Architecture:

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 512, 512, 3)]     0

 conv2d (Conv2D)             (None, 512, 512, 32)      896

 max_pooling2d (MaxPooling2D  (None, 256, 256, 32)     0
 )

 conv2d_1 (Conv2D)           (None, 256, 256, 64)      18496

 max_pooling2d_1 (MaxPooling  (None, 128, 128, 64)     0
 2D)

 flatten (Flatten)           (None, 1048576)           0

 dense (Dense)               (None, 100)               104857700

 dense_1 (Dense)             (None, 1)                 101

=================================================================
Total params: 104,877,193
Trainable params: 104,877,193
Non-trainable params: 0
```

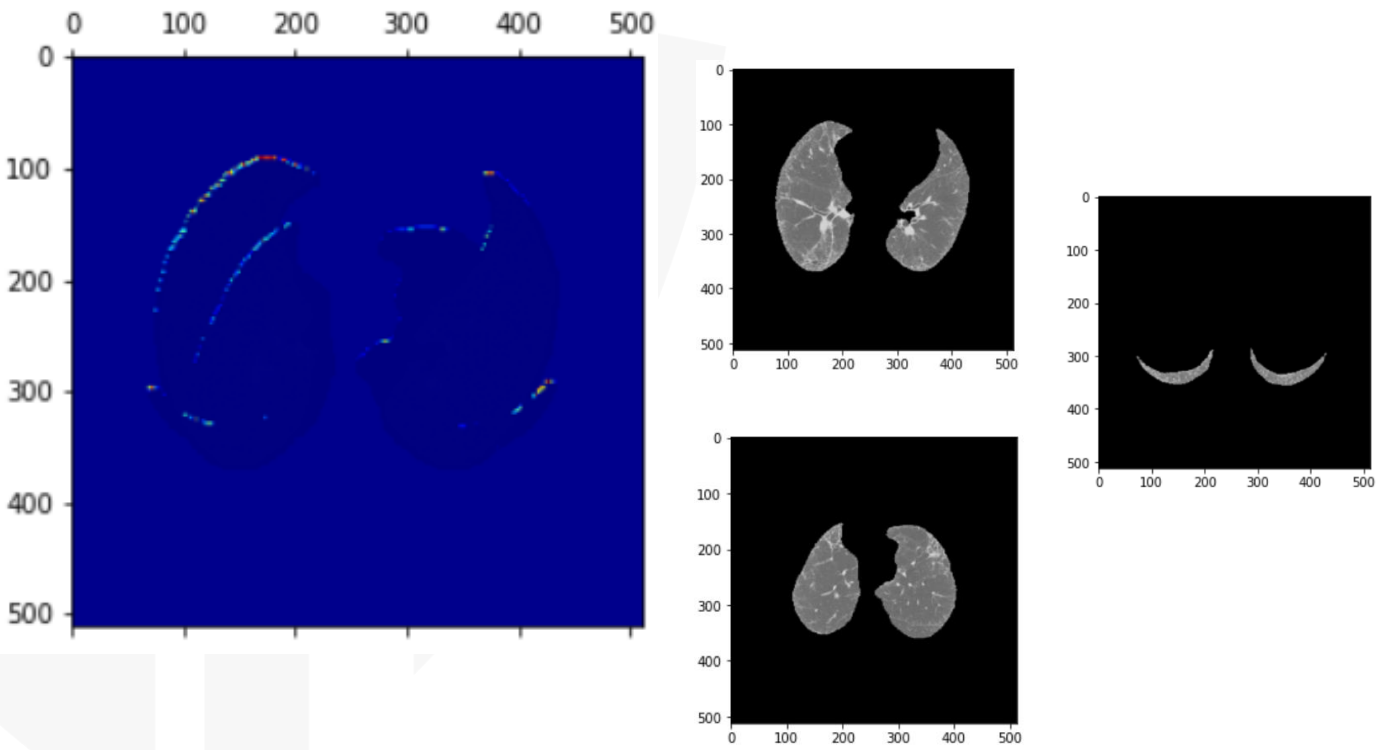
Stage 3 Overall Process:

- Get Lung Segmented CT Scan – Scan has n slices
- Scale each slice of each scan into 512x512 shape
- For each scan extract 3 slices – Highly representative of Top lobe , Middle Lobe and Bottom Lobe
- The Percentage health from patient demographic information is used to train the network
- Loss used to train network : MAE
- Extract the penultimate layer – 100 features per scan



Stage 3 – Encoding and Latent Space Features

- Analysing our Network – Grad CAM
  - Is it looking in the right places ?



Some Extracted Features of F0 to F99

F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
22.1083	0	23.6127	24.1468	0	20.5862	0	0	28.5925	0	32.0186	0
5.91016	0	32.1422	38.0218	0	26.1051	0	0	30.8499	0	36.3694	0
18.4812	0	23.6897	32.6728	0	20.1726	0	0	30.6637	0	35.3629	0
27.9803	0	33.7637	37.3129	0	27.2114	0	0	17.1845	20.5368	26.5282	0
35.8003	0	33.9044	41.964	0	35.603	0	0	39.855	12.7562	43.8762	0
21.5914	0	17.8634	22.3152	0	19.1859	0	0	26.8763	0	30.5447	0
10.0553	0	14.6571	24.679	0	13.4737	0	0	7.16965	6.34626	9.78482	0
30.5674	0	26.5914	36.0165	0	27.8781	0	0	33.3844	9.36808	36.9799	0
16.3665	0	13.0591	14.7006	0	17.2647	0	0	25.6001	0	25.9761	0
21.9901	0	25.4308	28.4972	0	24.3831	0	0	18.0172	9.37004	23.9744	0
20.2339	0	18.3128	22.1917	0	20.1326	0	0	17.4246	8.93783	23.4787	0
35.1441	0	31.364	43.6304	0	34.1605	0	0	33.4359	11.2458	41.0487	0
32.9572	0	32.7892	40.2274	0	31.9496	0	0	28.8771	14.8562	36.3797	0
28.3582	0	30.901	36.1566	0	29.6214	0	0	25.2661	17.2289	32.3779	0
18.0424	0	23.3639	21.4072	0	19.6519	0	0	25.7443	0	27.4231	0
32.3091	0	27.9843	0	0	26.1804	0	0	37.0183	18.2659	35.3493	0
25.7069	0	26.9894	30.0572	0	27.3298	0	0	22.1135	21.6876	28.2415	0
25.0366	0	32.0692	36.9576	0	26.6444	0	0	25.0399	8.79115	32.9985	0
20.3824	0	12.5754	34.8785	0	9.86813	0	0	29.9865	0	33.8439	0
35.4077	0	36.0411	49.9839	0	33.3334	0	0	51.413	0	54.7107	0
35.0451	0	39.6059	39.926	0	35.5696	0	0	40.2395	16.6263	44.2581	0
21.1941	0	24.3754	25.3424	0	23.171	0	0	22.3827	9.28525	25.1763	0
16.2468	0	18.228	21.9778	0	16.1668	0	0	12.2824	11.142	16.8253	0
15.6681	0	17.921	18.2285	0	10.7965	0	0	23.2862	0	25.0426	0
23.4306	0	29.9826	36.7238	0	31.809	0	0	29.1166	12.2491	33.7789	0
32.6512	0	27.4321	10.3016	0	31.2442	0	0	43.5685	23.9551	38.4889	0
25.694	0	22.8061	33.0264	0	19.076	0	0	33.1195	0	38.6029	0
17.6811	0	18.5621	16.7106	0	17.8901	0	0	16.7327	8.88593	17.8399	0
37.4137	0	34.5837	45.5256	0	39.9487	0	0	30.1776	22.7026	38.4824	0
23.5113	0	25.2304	34.3499	0	22.6669	0	0	25.7702	16.4367	29.6546	0
24.2192	0	24.175	21.7135	0	22.5665	0	0	26.162	0	29.2153	0
19.2206	0	11.9207	22.0417	0	12.2456	0	0	39.0232	0	38.9077	0
24.7719	0	24.9981	30.9783	0	24.2277	0	0	31.1294	13.049	34.1755	0
12.1586	0	11.3136	16.9495	0	15.8541	0	0	14.2716	0	16.8997	0
34.1907	0	37.8666	44.1074	0	38.5107	0	0	31.6226	16.1452	36.5353	0

Predicting FVC

- We combine all the three feature concepts – Demographics , handcrafted and latent features
- Totally we have 118 features to estimate FVC
- We built linear and polynomial regression models and compared performances based on RMSE loss

Results

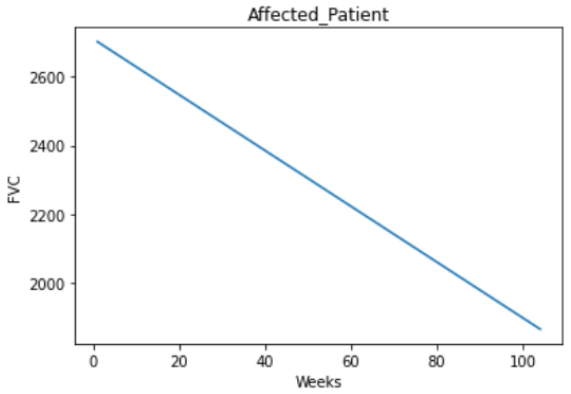
	Demographics only	Demographics + Handcrafted	Demographics + Handcrafted+Latent
Linear Regression - RMSE loss Train	368.9646	317.8707	245.7642
Linear Regression - RMSE loss Test	451.3221	380.8113	525.487
Polynomial Regression - degree 2 - RMSE loss Train	316.0495	14.9648	10.1282
Polynomial Regression - degree 2 - RMSE loss Test	365.312	9094.1894	7050730495

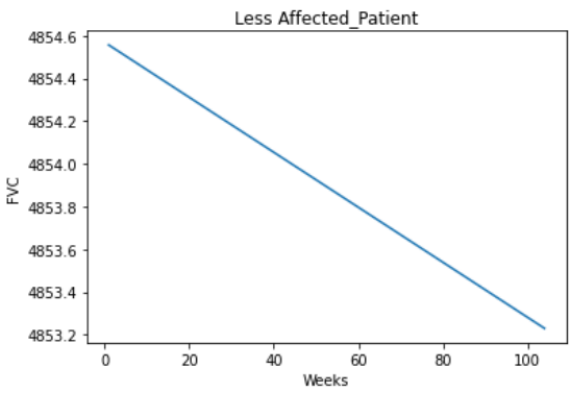
Observations

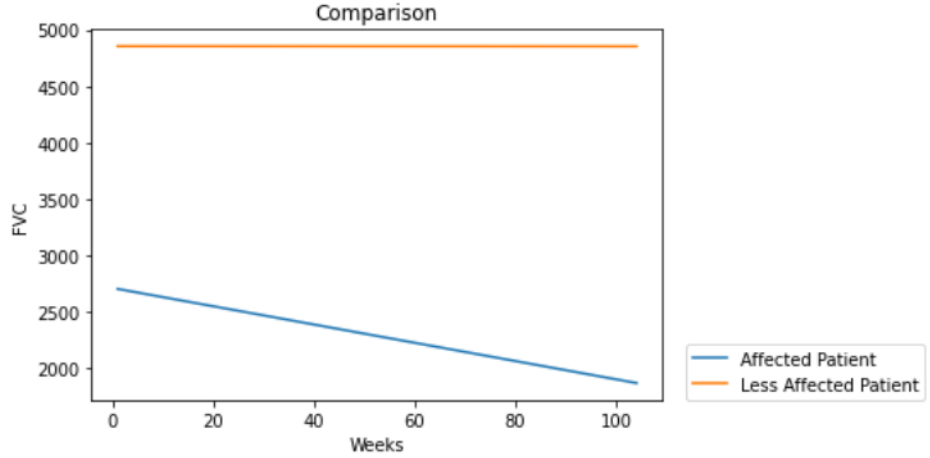
- The polynomial model seems to be over fitting and also with latent features we see overfitting generally
- Overall loss looks not that great but when analysed against true value we were able to understand the behaviour. The true FVC is highly fluctuating and that resulted in higher RMSE value

Visual Results – Discussion and Conclusion

- We took two patients – Highly affected and less affected and predicted FVC for 104 weeks
- Both show decline in lung function – FVC decreases over time
- For highly affected patient the rate of decline is faster – The model seems to be predicting the trend properly







Conclusion and Future Works:

- The model seems to be good in predicting the overall trend in decline but is not great in predicting the exact FVC value
- In future, if we allow FVC's initial seed values as part of feature matrix – the rmse might reduce
- Also if we have a memory based networks we might be able to capture non-linear trends better.