

COMPUTER AIDED DIAGNOSTIC SYSTEM FOR LUNG DISEASES

A Project Report

Submitted by

**HARSH AJMERA
NAMAN BANSAL
ABHISHEK JAIN
ADITYA MITTAL**

Under the Guidance of

Prof Mrs. Supriya Agarwal

*in partial fulfillment
for the award of the degree of*

**MBA Tech.
COMPUTER ENGINEERING
At**



**MUKESH PATEL SCHOOL OF TECHNOLOGY
MANAGEMENT & ENGINEERING
April 2021**

DECLARATION

We, Harsh Ajmera(N206), Naman Bansal(N211), Abhishek Jain(N231) and Aditya Mittal(N250), MBA Tech. (Computer Engineering), VIII semester understand that plagiarism is defined as anyone or combination of the following:

1. Un-credited verbatim copying of individual sentences, paragraphs or illustration (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.
2. Un-credited improper paraphrasing of pages paragraphs (changing a few words phrases, or rearranging the original sentence order)
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who wrote what. (Source: IEEE, The institute, Dec. 2004)
4. We have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of our work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.
5. We affirm that no portion of our work can be considered as plagiarism and we take full responsibility if such a complaint occurs. We understand fully well that the guide of the seminar/ project report may not be able to check for the possibility of such incidents of plagiarism in this body of work.

HARSH AJMERA (N206)

NAMAN BANSAL (N211)

ABHISHEK JAIN (N231)

ADITYA MITTAL (N250)

Place: Mumbai

Date:

CERTIFICATE

This is to certify that the project entitled Speech Emotion Recognition Model is the bonafide work carried out by Harsh Ajmera(N206), Naman Bansal(N211), Abhishek Jain(N231) and Aditya Mittal(N250) of MBA Tech. (Computer Engineering), MPSTME (NMIMS), Mumbai, during the VIII semester of the academic year 2020-21, in partial fulfillment of the requirements for the award of the Degree of MBA in Technology Management as per the norms prescribed by NMIMS. The project work has been assessed and found to be satisfactory.

Prof Supriya Agarwal

Internal Mentor

Examiner 1

Examiner 2

Dean

Acknowledgment

This project would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them.

We are highly indebted to Mukesh Patel School of Technology Management and Engineering, Narsee Monjee Institute of Management Studies for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

We would like to express my gratitude towards our parents and friends. We would especially like to devote a vote of thanks to Prof Supriya Agarwal for her kind cooperation and encouragement which helped us in completing this phase of the project. We also would like to thank her for showing us some examples that related to the topic of our project.

We would like to express our special gratitude and thanks to Prof. Abhay Kolhe for giving us such attention and time. He has encouraged and guided us through several meetings.

Our thanks and appreciations also go to our colleague in developing the project and people who have willingly helped us out with their abilities. We perceive this opportunity as a big milestone in our career development. WE will strive to use gained skills and knowledge in the best possible way, and we will continue to work on their improvement, in order to attain desired career objectives.

HARSH AJMERA (N206)

NAMAN BANSAL (N211)

ABHISHEK JAIN (N231)

ADITYA MITTAL (N250)

Place: Mumbai

Date:

TABLE OF CONTENTS

ABBREVIATIONS	vii
ABSTRACT.....	viii
1. Introduction.....	Error! Bookmark not defined.
1.1 AIM.....	ix
1.2PURPOSE.....	x
2. Review of Literature	Error! Bookmark not defined.
2.1 Art’s First Place Solution.....	xi
2.2 Abhishek Bhat’s QuantileReg + Linear Decay solution:	xii
2.3 Dr. Konya’s domain expert insights	xiii
2.4 Boxiang Yun’s solution using XGBoost.....	xiii
2.5 Thinking beyond the mean: a practical guide for using quantile regression methods for health services research.	xiv
2.6 EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks	xiv
2.7 Learning to forget: continual prediction with LSTM.....	xv
2.8 Rohan Rao’s Understanding Laplace Log Likelihood Notebook	xvi
2.9 Yasufumi Nakama’s LGB baseline approach:	xvii
3. Analysis & Design	xix
3.1 ATTRIBUTES IN THE DATASET	xix
3.3 Handling the CT scan data.	xx
3.4 Handling of the statistical data.....	xxiii
3.5 Modeling the data	xxvi
3.5.2 What is Laplace Log Likelihood?	xxvii
3.6 The LSTM Layer	xxix
3.7 The Dropout Layer.....	xxx
3.8 Training the model.	Error! Bookmark not defined.
4. Implementation	xxxiv
4.1 CAD System	xxxiv
5. RESULTS AND DECLARATION	xxxvi
5.1 Conclusion and Future Work	xxxvi
REFERENCES	xxxviii

TABLE OF FIGURES

Figure 1 Standard OLS methods (left) V/S Quantile Regression(right)	xiv
Figure 2 Model Scaling for Convolutional Neural Networks	xv
Figure 3 LSTM model with combined Forget Gate.....	xvi
Figure 4 Percent attribute relations	xx
Figure 5 Segregating the DICOM Dataset.....	xx
Figure 6 Step by Step implementation to apply mask of the CT scan	xxi
Figure 7 Steps involved to extract Statistical Data of DICOM files	xxiii
Figure 8 Conversion of data (Baseline Approach).....	xxiv
Figure 9 Attributes Description	xxv
Figure 10 PEARSON Correlation between Extracted Features.....	xxvi
Figure 11 Graphical Representation to understand Quantile Regression	xxvii
Figure 12 The journey of the data in our solution.....	xxviii
Figure 13 LSTM Model Implementation in our model	xxix
Figure 14 The Droupout Layer	xxx
Figure 15 Model Training.....	xxxii
Figure 16 Training score progression of each fold	xxxiii
Figure 17 The doctor will fill the above form and upload the CT scan of the patient.	xxxiv
Figure 18 The process flow of the web application	xxxv
Figure 19 Once the CT scan files gets uploaded, the doctor will be able to see the prediction graph.	xxxv

ABBREVIATIONS

NN – Neural Network

CNN – Convolutional Neural Network

RNN – Recurrent Neural Network

FVC – Forced Vital Capacity

ILD – Interstitial Lung Disease

OSIC – Open-Source Imaging Consortium

DL – Deep Learning

LSTM – Long Short-Term Memory

DFT – Discrete Fourier Transform

CT Scan - Computerized Tomography Scan

KPI – Key Performance Indicator

DICOM - Digital Imaging and Communications in Medicine

ABSTRACT

In this project, we use the power of modern deep neural networks to solve a severe healthcare problem that has been troubling both doctors and patients for years, the problem of pulmonary fibrosis diagnosis. Pulmonary Fibrosis is a lung disease that occurs when lung tissue becomes damaged and scarred, but the fact that makes it the one of the most troubling disease is that it is very hard to diagnose. Even experienced doctors have trouble figuring out the proper diagnosis of this disease because the severity of this disease fluctuates very rapidly. Patients can go from experiencing mild symptoms to suffering harsh symptoms in a matter of weeks.

With the help of modern learning technologies, we are sure we can build a system that can help doctors predict the future course of this disease so that they can prescribe the best treatment to their patients. We harness the power of Recurrent Neural Networks (RNNs) that can understand relationships in data over time. We will be using the Long Short-Term Memory RNN (LSTM) to predict a Forced Vital Capacity score (FVC score), which is, basically, a health performance measure of the lungs.

Thus, we aim to build a deep neural network that can predict the FVC score of the patient. We hope that our solution will someday actually be used by doctors to provide the best treatment to pulmonary fibrosis patients.

1. INTRODUCTION

Pulmonary Fibrosis is a terminal interstitial lung disease where the patient faces severe breathing problems such as shortness of breath (dyspnea) and inflammation of lungs. It occurs when the lung tissue gets irreversibly damaged and scarred which leads to thickening of the lung tissue, thereby, causing breathing problems.

Even though the damage caused by pulmonary fibrosis cannot be reversed, proper medication and therapies can sometimes bring comfort by easing symptoms and improving the quality of life. But given the uncertain idiopathic nature of the disease, it is difficult for doctors to suggest the best treatment. This is what makes pulmonary fibrosis one of the most troubling diseases; patient's condition may shift from seemingly long-term stable to rapid deterioration in a very short span of time. What makes it even worse is the fact that, current methods make the disease difficult to treat, even with access to a chest CT scan. [20][21]

Apart from the current diagnosis methods, the lack of awareness of the uncertain nature of pulmonary fibrosis kept modern prediction methods of machine learning and deep learning away from this domain. Thanks to the efforts of Open-Source Imaging Consortium (OSIC) the pulmonary fibrosis prediction problem was introduced to the machine learning world when they put together a dataset of 175 pulmonary fibrosis patients from different backgrounds and organized a competition on Kaggle with an award of \$55,000.

In this project, with the help of deep neural networks we want to predict the future course of the disease to help their doctors provide them the most optimal treatment. We will thus perform a time series prediction to estimate the FVC of the patient, which, is a health KPI of the lungs that fluctuates and deteriorates over time during this disease. [22][23]

1.1 AIM

In this project, we aim to predict the FVC score for the next 50 weeks for the patient with the help of simple attributes such as age, sex, smoking status, initial FVC and the CT scan.

We also aim to build a full stack web application where doctors can upload the patient's details easily and in no-time get the future FVC predictions for the patient.

1.2 PURPOSE

Using this project, we are trying to nullify the diagnostic error and provide another set of diagnostic tools which will help the doctor in better treating the patient with data driven diagnosis.

Thus, we want to equip pulmonary fibrosis doctors with a reliable friend that will give them a better idea of the patient's condition.

2. REVIEW OF LITERATURE

2.1 Art's First Place Solution

Art's solution was a result of blend of 2 models: EfficientNet and Quantile Regression based LSTM network. The EfficientNet's B5 worked on the CT scan slices, whereas the LSTM network was used to make final FVC predictions from the statistical tabular data. The author then manually blended the predictions of both the models to get the final prediction.

Here are the exact steps on how Art achieved the 1st place on the private leaderboard.

1. Trained both models from scratch. For EfficientNet B5 with 30 epochs and for Quantile Regression 600 epochs to train.
2. Did some feature filtering, by removing the precomputed "Percent" feature which made the predictions worse (probably because this feature was precomputed).
3. In terms of model blending, gave a higher score to the Quantile Regression model, because from my point of view it was more reliable.

Advantages of Art's method:

1. Uses the power of ensemble learning, by blending results of EfficientNet and LSTM network.
2. The LSTM network converges very quickly.

Disadvantages of Art's method:

1. Manually blends both the models to adjust predictions for the best possible outcome.
2. Ignores data augmentation methods available.
3. Ignores data from DICOM files such as lung volume, area, etc., which have a decent correlation with the FVC value.

Link to Art's First place solution:

<https://www.kaggle.com/artkulak/inference-45-55-600-epochs-tuned-effnet-b5-30-ep>

2.2 Abhishek Bhat's QuantileReg + Linear Decay solution

Abhishek Bhat's solution uses similar models as that of Art. Quantile Regression based LSTM network ensembled with a variant of EfficientNet. The author uses EfficientNet to calculate the linear decay in the lung tissue. The EfficientNet CNN is also given data like average lung tissue per cm³, average lung tissue per thickness and average total lung tissue. Ther predictions made by the EfficientNet were blended with the predictions of the LSTM networks that worked only with the statistical data.

The author finally uses K=fold method, with K=10 to train both the networks on their respective data before simply blending the best predictions.

Advantages of Abhishek Bhat's solution:

Several features like Lung Volume, Average tissue area for each patient etc. were extracted from the CT scans.

Average tissue area was found to show a high positive correlation with the decline in FVC over time, hence a couple of features related to tissue were used in the model.

Uses K-fold method to eliminate training data irregularities.

Disadvantages of Abhishek Bhat's solution:

The solution becomes too big and too complex.

Huge amount of training data.

Focuses more on the use of EfficientNet CNN for the prediction, ignoring the fact that the other network is a LSTM RNN network which can store short term temporal relationships in data.

Link: <https://www.kaggle.com/abhishekgbhat/quantreg-linear-decay-efficientnet-b1-su#Quantile-Regression>

2.3 Dr. Konya's domain expert insights

Dr. Konya helped us in understanding the depth of the problem statement and established our basic understanding of medical parameters and terminologies such as Forced Vital Capacity (FVC), Hounsfield units (HU), factors affecting idiopathic nature etc.

The author also takes a deep dive into the world of CT scans by explaining us the various important details a DICOM files stores and teaches us how to segment lung tissue from the CT scan images. The author also explains how parameters such as mean, median and kurtosis of the CT scan image have a high correlation with FVC score predictions.

The author has two parts of the notebook and both were very insightful, even though, it did not present us with the solution, the notebooks are aimed towards pointing us in the direction of the solution.

Link to Dr. Konya's domain expert insights:

<https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/discussion/165727>

2.4 Boxiang Yun's solution using XGBoost

The author works on Dr. Konya's guidelines towards the solution to the problem. The author in his notebook focuses on using the CT images currently and getting the most out of them. The author extracts feature like lung volume, tissue areas and Hounsfield's histogram's mean, skew and kurtosis before training a XGBoost model on the data.

The author also calculates different Hounsfield ratios for the depth of the CT scans and then trains the XGBoost model with custom early stopping function and k-fold where $k=5$.

Link to Boxiang Yun's solution:

<https://www.kaggle.com/hfutybx/osic-feature-extract-from-ct>

2.5 Thinking beyond the mean: a practical guide for using quantile regression methods for health services research.

This is a paper which discussed how modelling methods that are based around the mean of the data fail to perform when it comes to health applications. The paper addresses the fact that healthcare or disease data is majorly skewed towards a certain type of people with similar attributes that usually face the disease. The data has very few outliers, but the outliers are the ones that face the most sever symptoms.

The authors focus on the fact that quantile regression is a median based method which allow analysis to move away from the mean and see median as an alternative to least squares regression and related methods, which typically assume that the associations between independent and dependent variables are all at similar levels.

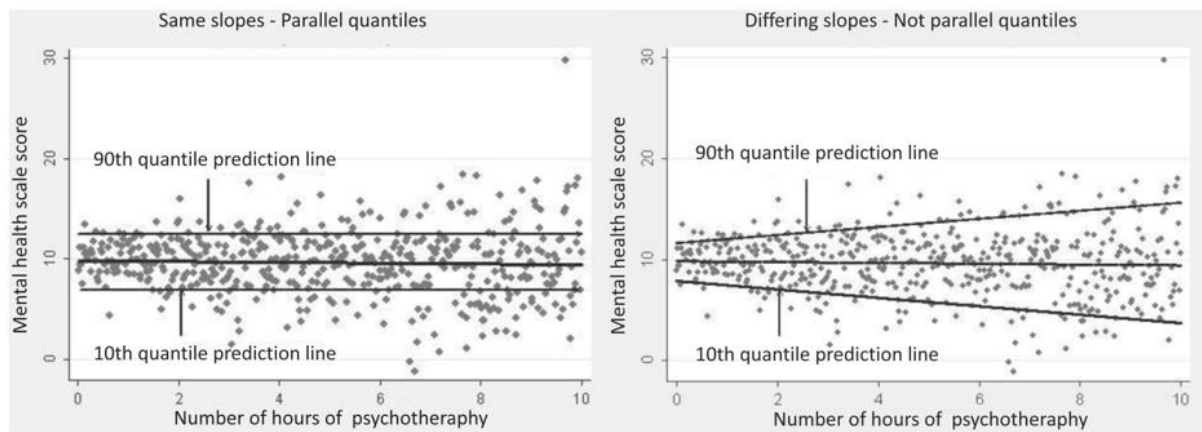


Figure 1 Standard OLS methods (left) V/S Quantile Regression(right)

Link:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4054530/#:~:text=The%20main%20advantage%20of%20quantile,nonlinear%20relationships%20with%20predictor%20variables.>

2.6 EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

DOI: arXiv:1905.11946

This paper introduced us to EfficientNets, what makes them better, and why should we use them. The days of the AlexNet or Google's InceptionNets are over. EfficientNets are a new type of CNN, specially introduced by Google AI Labs, that are built to scale.

EfficientNets achieve much better accuracy and efficiency than previous ConvNets. The EfficientNet-B7 variant achieves state-of-the-art 84.3% top-1 accuracy on ImageNet, while being 8.4x smaller and 6.1x faster on inference than the best existing ConvNet. Our EfficientNets also transfer well and achieve state-of-the-art accuracy on CIFAR-100 (91.7%), Flowers (98.8%), and 3 other transfer learning datasets, with an order of magnitude fewer

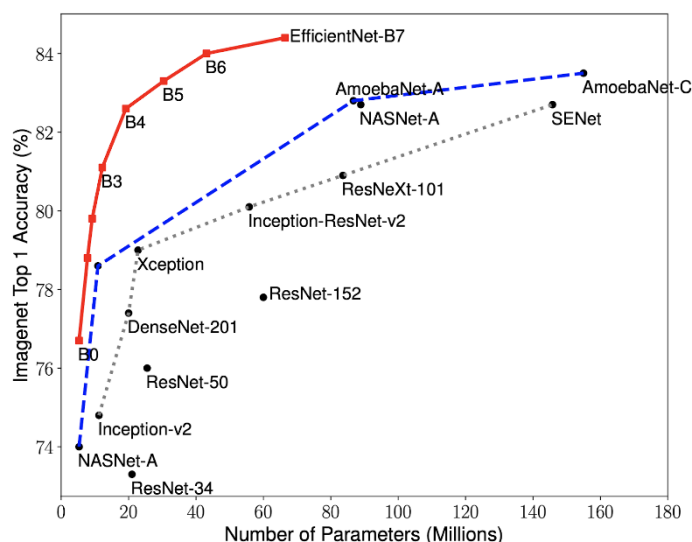


Figure 2 Model Scaling for Convolutional Neural Networks

parameters.

Link: <https://arxiv.org/abs/1905.11946v5>

2.7 Learning to forget: continual prediction with LSTM

DOI: 10.1049/cp:19991218

In this paper, the author introduces us to the essence of Long short-term memory networks, a RNN, which has the ability to store long and short term temporal relationships between data. The paper also talks about the applications of LSTMs and what makes them useful. The paper also discusses LSTM's flaws and how a LSTM network with a combined forget gate can solve those problems (just like the name suggests).

The paper was helpful to us in understanding applications of LSTM networks to our problem statement along with its shortcomings.

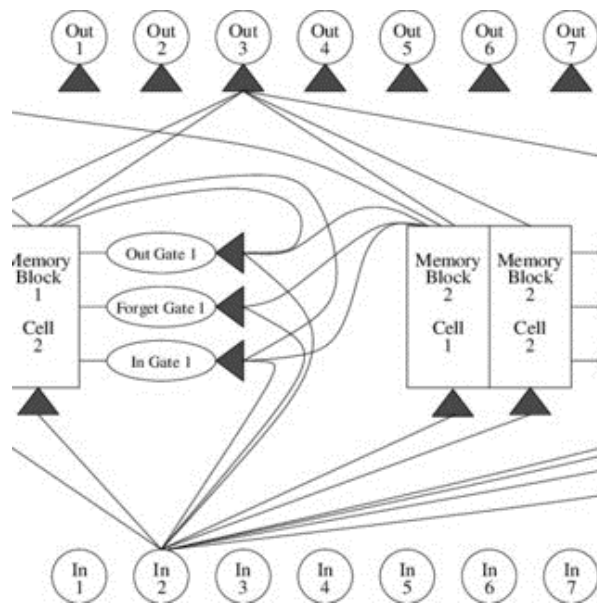


Figure 3 LSTM model with combined Forget Gate

2.8 Rohan Rao's Understanding Laplace Log Likelihood Notebook

In this notebook, the author takes a deeper dive into the evaluation metric given by Kaggle to score the submissions. Understanding the Laplace Log Likelihood metric was crucial to us as it was the only true quantitative measure to compare our solutions with that of others. This notebook helped us understand a great deal about the behavior of the metric and its relations with its parameters.

$$\sigma_{clipped} = \max(\sigma, 70),$$

$$\Delta = \min(|FVC_{true} - FVC_{predicted}|, 1000),$$

$$metric = -\frac{\sqrt{2}\Delta}{\sigma_{clipped}} - \ln(\sqrt{2}\sigma_{clipped}).$$

In this notebook, the author takes the mean of the given FVC values as predictions for the same values and then calculates the metric. Let's say, for a patient in the training set we have a set of 10 FVC reading, and we simply take the average of these FVC reading and project it as the prediction in the Laplace log likelihood formula and let's say it returns us a value of -8.61. This means that our model needs to have a LLL score of more than -8.61 or else it is just obsolete.

Link: <https://www.kaggle.com/rohanrao/osic-understanding-laplace-log-likelihood#Constant-Prediction>

2.9 Yasufumi Nakama's LGB baseline approach

In this notebook, the author uses a smart data augmentation technique to use data to the model's advantage; the author (& the rest of the participants) call this approach the 'baseline approach'.

To describe the baseline approach in short, for a patient, let's say we have FVC observation for 10 different weeks. We keep the 0th observation as base week and then use the rest of the weeks one-by-one as the final week, and the FVC value of the final week as the value-to-be-predicted for the final week.

The author applied the data augmentation technique to multiply the data rows and then used a LGBM model with K-fold and early stopping to train the model.

Link: <https://www.kaggle.com/yasufuminakama/osic-lgb-baseline>

3. ANALYSIS & DESIGN

Our solution for the problem statement is a mixture of various solution posted by participants on Kaggle. We did our best to combine the strengths of each solution into our work and have come up with a quantile regression-based LSTM model. The model uses the baseline approach to augment the data and couples it with a variant of EfficientNet to handle the image data. We also extract parameters such as lung volume, lung area, window width, pixel spacing etc, from the CT scans.

3.1 ATTRIBUTES IN THE DATASET

For this project, we took the dataset from the OSIC Website which had a Dataset Available for this. This Dataset file contains data of 175 different patients which we have used in our model.

1. ID of patient (Used to track data of Individual Patient)
2. Week Number
3. Week Number is noted to track the patient's lung status via FVC score check.
4. Week – 0 is when the CT scan of the patient takes place.
5. FVC - Forced vital capacity is the amount of air that can be forcibly exhaled from your lungs after taking the deepest breath possible, as measured by spirometry.
6. Percent – Percent is converted via FVC score which tells the status of Lung health.

Percent Value	Relation with a normal person
80% or greater	Normal
70%–79%	Mildly abnormal
60%–69%	Moderately abnormal
50%–59%	Moderate to severely abnormal
35%–49%	Severely abnormal

Less than 35%	Very severely abnormal
---------------	------------------------

Figure 4 Percent attribute relations

3.2 HANDLING THE CT SCAN DATA

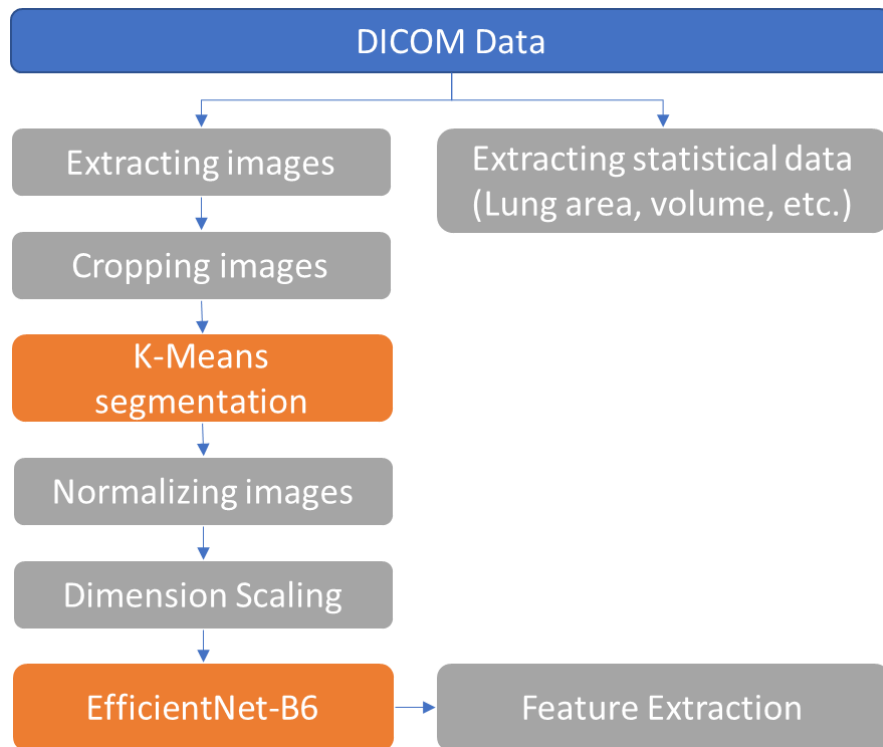


Figure 5 Segregating the DICOM Dataset

Every patient has multiple DICOM files. But the attributes like the window length, window width, pixel spacing etc. are the same inside each DICOM file except for the image data inside the DICOM files.

Therefore, we first extract the statistical data from one of the DICOM files. Then, we move on to extract image data from the DICOM files. For each image, we crop it to 512x512pixels, apply K-Means segmentation (where k=2) to segment the lung tissue from the other elements

present in the CT scan such as air, bone, etc. We then use erosion and dilation which has the net effect of removing tiny features like pulmonary noise. Using bounding boxes for each image label to identify which ones represent lung and which ones represent "everything else". We create the masks of the segmented lung and multiply it with the original image. [10]

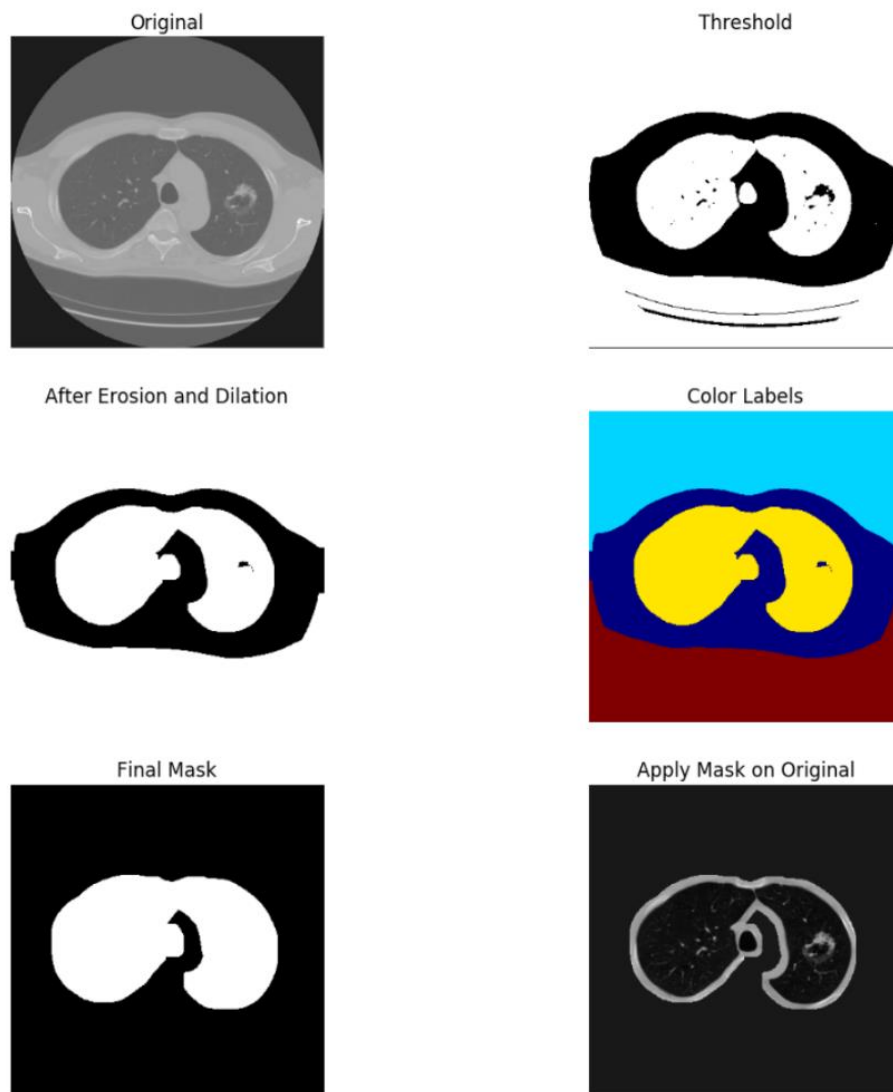


Figure 6 Step by Step implementation to apply mask of the CT scan

Once we have segmented the lung tissue from the image data extracted from the CT scans, we normalize the images so that every pixel is between 0 to 1.

We then merge the cropped, segmented, and normalized images from the CT scans into one variable and add a channel to it, this step is called dimension scaling. Now we are ready to extract features from the images using the EfficientNet-B6 model. We chose the B6 variant of the EfficientNet as the vector returned by it had the highest correlation from any other variant.

We combine then combine the vector given by EfficientNet and the statistical data extracted from the CT scan together and add this data with the statistical data from the dataset.

3.4 HANDLING THE STATISTICAL DATA

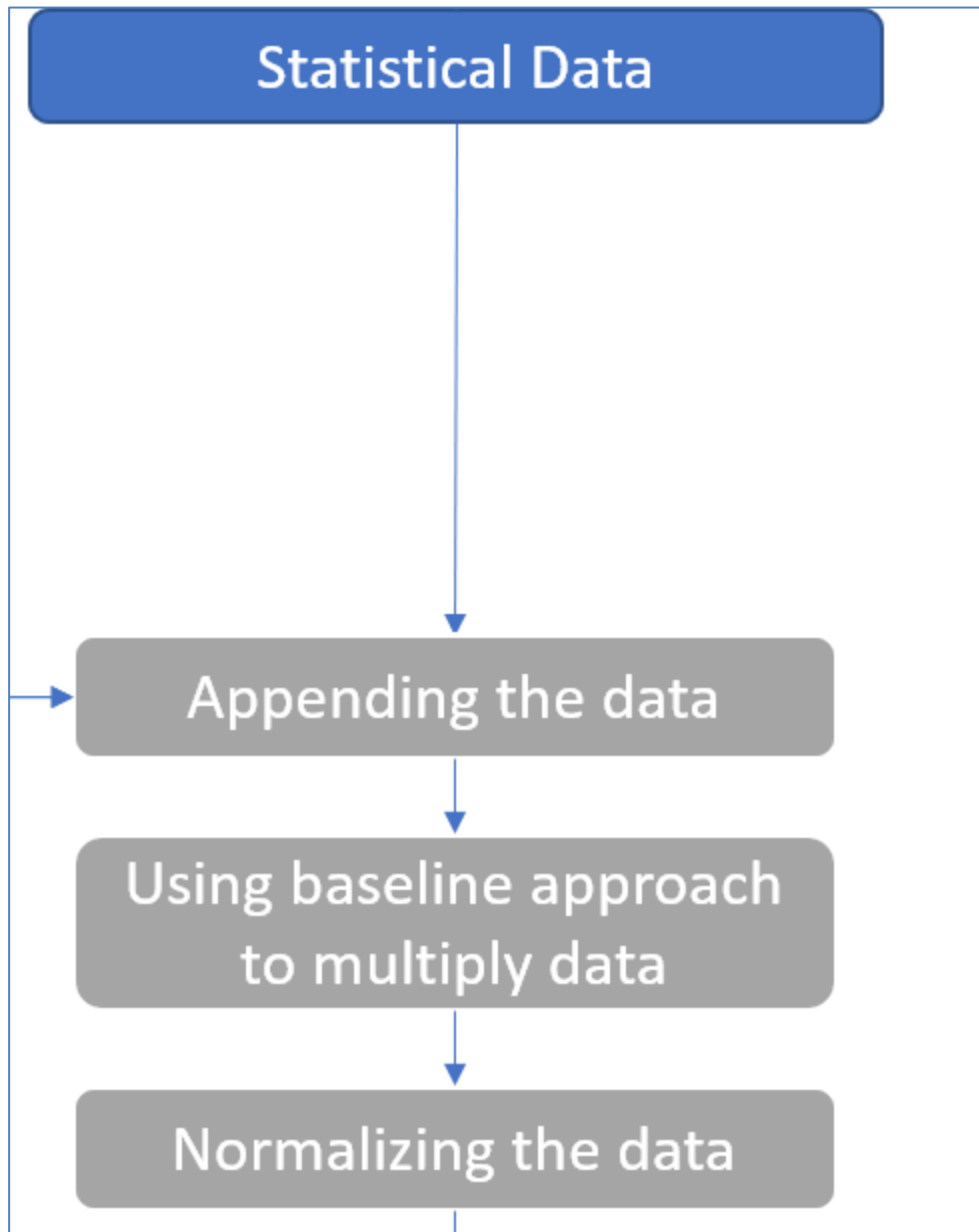


Figure 7 Steps involved to extract Statistical Data of DICOM files

After converting and appending the CT scan data to the dataset, we use the baseline approach to augment the data. As mentioned in Yasufumi Nakama's work, we change the structure of the data so that we can multiply the training data rows for our model.

To describe the baseline approach in short, for a patient, let's say we have FVC observation

for 4 different weeks. We keep the 0th observation as base week and then use the rest of the weeks one-by-one as the final week, and the FVC value of the final week as the value-to-be-predicted for the final week. Refer to the figure below for a tabular view of the same.

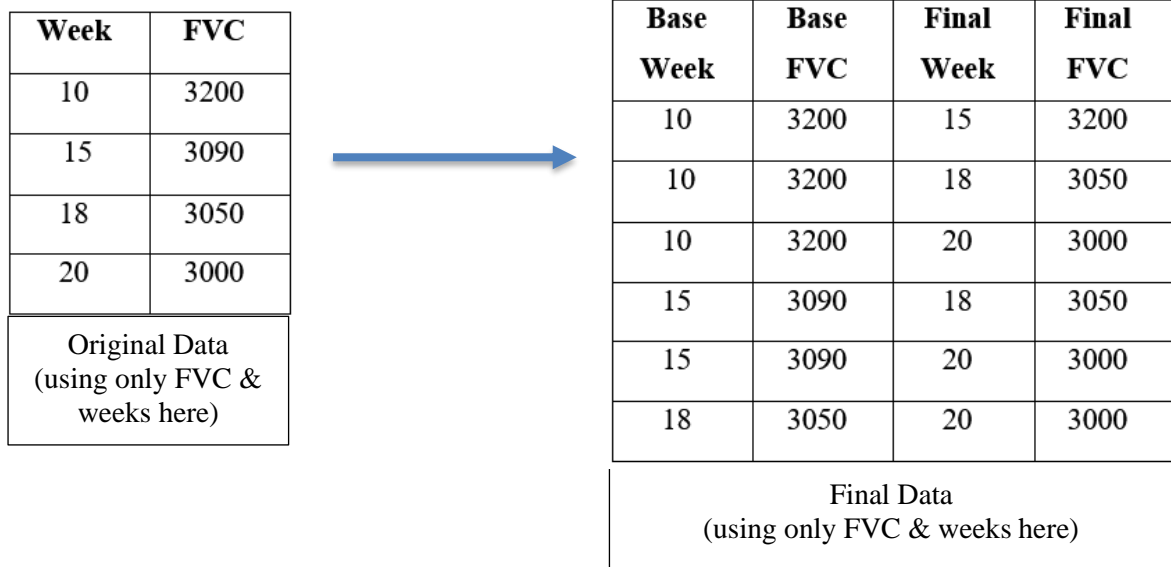


Figure 8 Conversion of data (Baseline Approach)

Then we normalize each feature, converting all values from 0 to 1.

Before moving forward, let us summarize the data data features have:

DATA FEATURE	DESCRIPTION
BASE WEEK	The base week of the observation
BASE PERCENT	The percentage FVC of the patient compared to a healthy person
BASE FVC	The base FVC of the observation
CURRENT WEEK	The current week being discussed in the observation
DIFFERENCE	The difference of week between the base week and current week
AGE	Age of the patient

SEX	Gender of the patient
SMOKING STATUS	The smoking status of the patient
AREA	The area of lung tissue in pixels
PIXEL SPACING ROW	The distance between two pixels in a row
PIXEL SPACING COLUMN	The distance between two pixels in a column
SLICE THICKNESS	The voxel depth of a slice
WINDOW WIDTH	The length of the slice
WINDOW LENGTH	The width of the slice
ROW DISTANCE	The length of the voxel
COLUMN DISTANCE	The width of the voxel
AREA IN CM2	The area of the lung tissue in cm2
VOLUME IN CM3	The volume of the lung tissue in cm3
EFFICIENTNET FEATURES	The vector returned by the EfficientNet
FINAL FVC	The FVC value of that week

Figure 9 Attributes Description

The Pearson's correlation of the datapoints is given below:

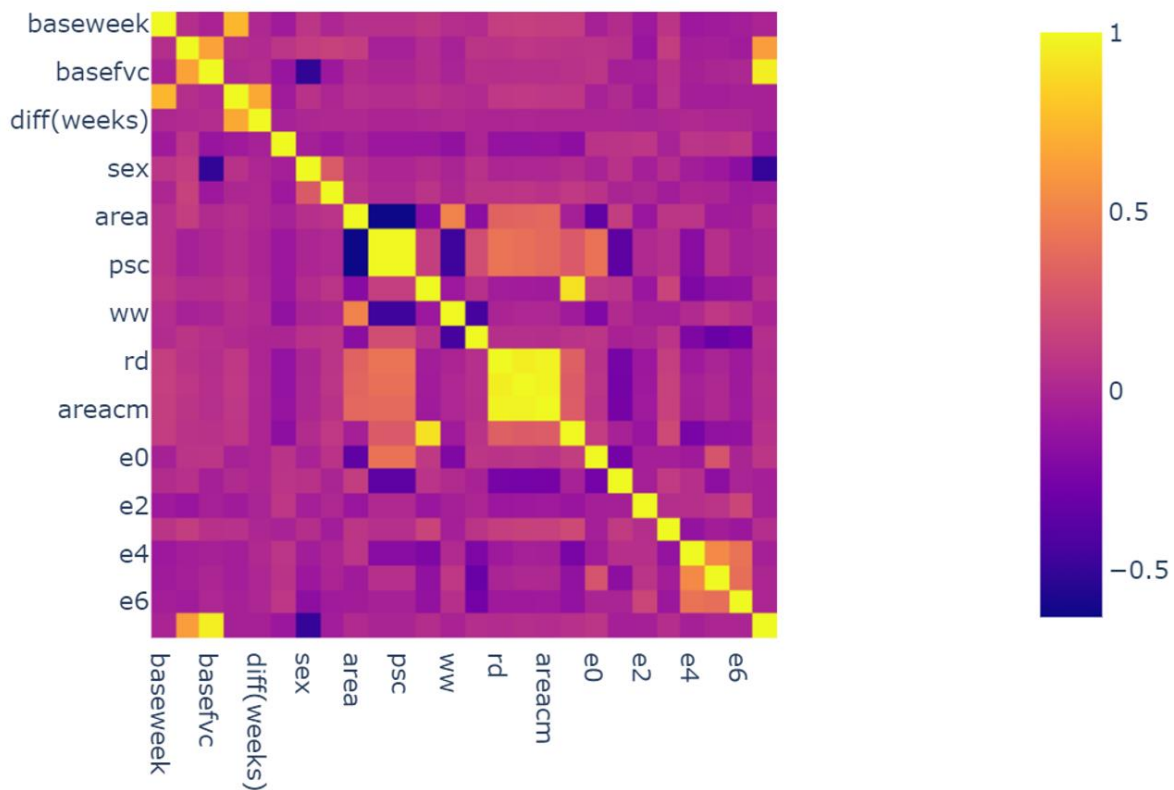


Figure 10 PEARSON Correlation between Extracted Features

In the above diagram, factors such as base FVC, base percent, sex, e0 and e3 have a better correlation than the rest of the data.

3.5 MODELLING THE DATA

As mentioned earlier, in this model we use a quantile regression-based LSTM network with a custom loss. To understand the model, let's break down and understand the major parts of the model such as quantile regression, Laplace log likelihood and the LSTM layers.

3.5.1 What is Quantile Regression?

Quantile regression is a median based method which allow analysis to move away from the mean and see median as an alternative to least squares regression and related methods, which typically assume that the associations between independent and dependent variables are all at similar levels.

In the figure below, we fit a OLS based regression line and three quantile regression-based regression (0.2, 0.5 and 0.8) on FVC vs week data graphs to give you a better idea about how quantile regression is better for capturing outliers.[24][25]

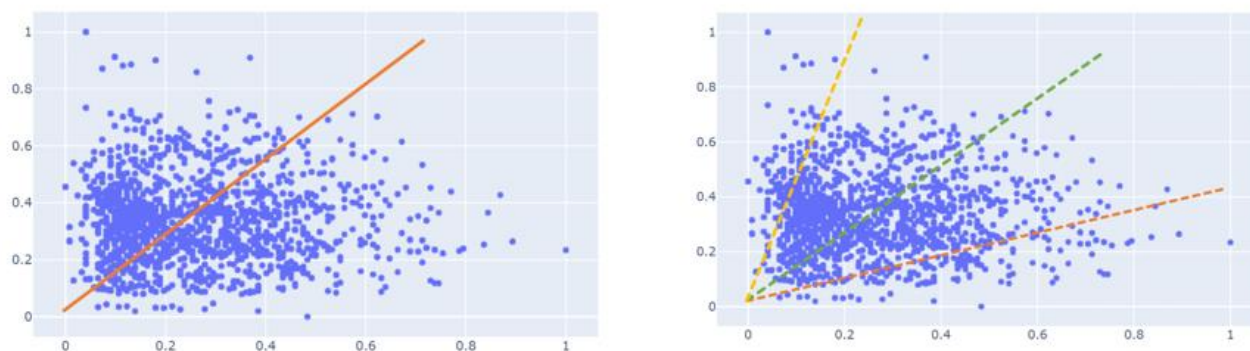


Figure 11 Graphical Representation to understand Quantile Regression.

3.5.2 What is Laplace Log Likelihood?

[11] A critical difference between probability and likelihood is in the interpretation of what is fixed and what can vary. In the case of a conditional probability, $P(D|H)$, the hypothesis is fixed, and the data are free to vary. Likelihood, however, is the opposite. The likelihood of a hypothesis, $L(H)$, is conditioned on the data, as if they are fixed while the hypothesis can vary. The distinction is subtle, so it is worth repeating: For conditional probability, the hypothesis is treated as a given, and the data are free to vary. For likelihood, the data are treated as a given, and the hypothesis varies.

For each true FVC measurement, you will predict both an FVC and a confidence measure. The metric is computed as:

$$\begin{aligned}\sigma_{clipped} &= \max(\sigma, 70), \\ \Delta &= \min(|FVC_{true} - FVC_{predicted}|, 1000), \\ metric &= -\frac{\sqrt{2}\Delta}{\sigma_{clipped}} - \ln(\sqrt{2}\sigma_{clipped}).\end{aligned}$$

The error is thresholds at 1000 to avoid large errors adversely penalizing results, while the confidence values are clipped at 70 ml to reflect the approximate measurement uncertainty in FVC. The final score is calculated by averaging the metric across all the observations.[26]

Before moving to the details of the model, let us summarize the entire journey of the data with the below figure.

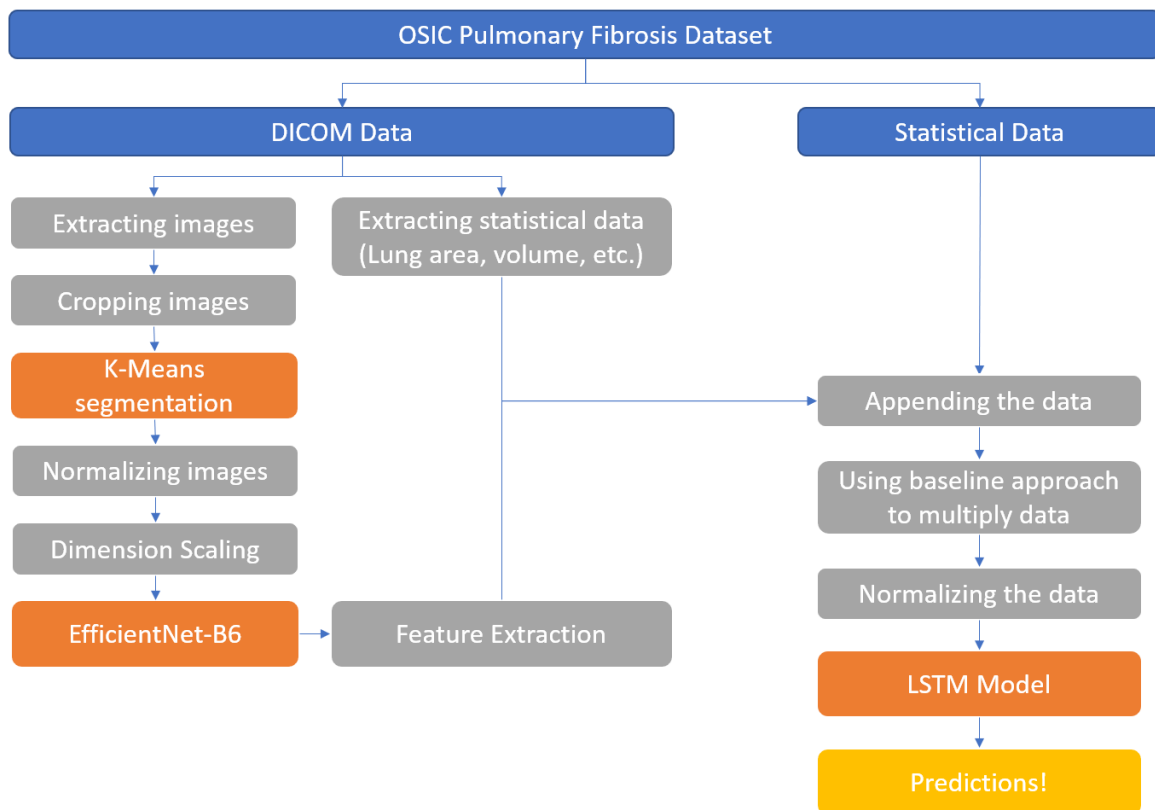


Figure 12 The journey of the data in our solution

Now that we have a clear understanding of how data is travelling, let us take a closer look at the LSTM model.

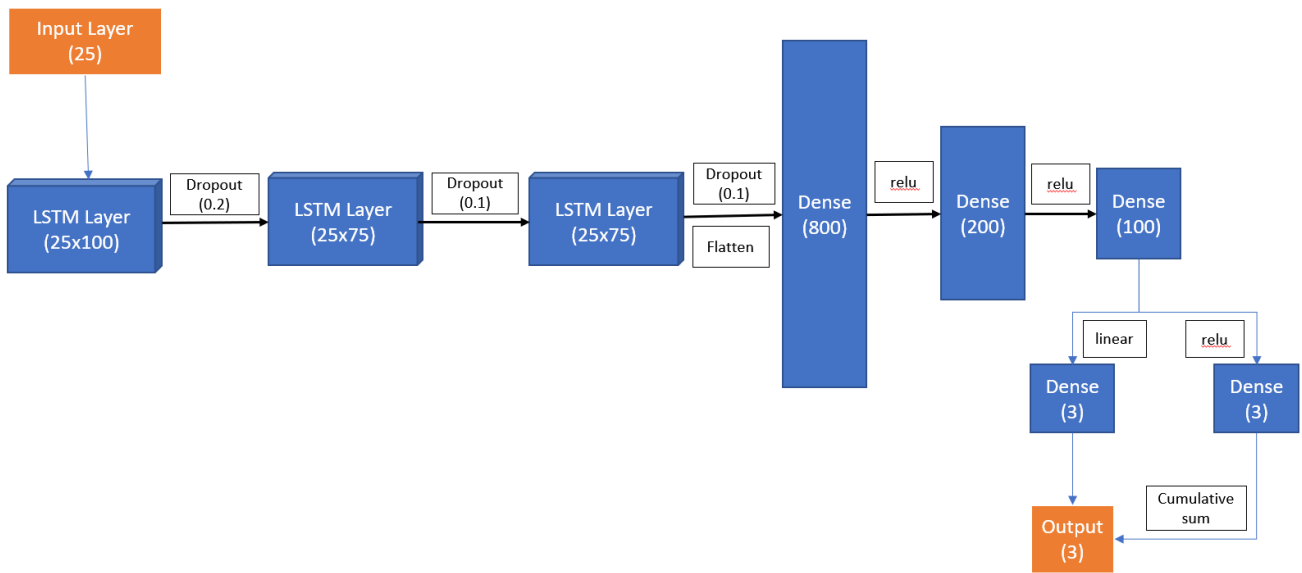


Figure 13 LSTM Model Implementation in our model

3.6 THE LSTM LAYER

LSTM has a special architecture which enables it to forget the unnecessary information. The sigmoid layer takes the input $X(t)$ and $h(t-1)$ and decides which parts from old output should be removed (by outputting a 0). This gate is called forget gate $f(t)$. The output of this gate is $f(t)*c(t-1)$.

The next step is to decide and store information from the new input $X(t)$ in the cell state. A Sigmoid layer decides which of the new information should be updated or ignored. A tanh layer creates a vector of all the possible values from the new input. These two are multiplied to update the new cell state. This new memory is then added to old memory $c(t-1)$ to give $c(t)$.

- Finally, we need to decide what we're going to output. A sigmoid layer decides which parts of the cell state we are going to output. Then, we put the cell state through a tanh generating all the possible values and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

3.7 THE DROPOUT LAYER

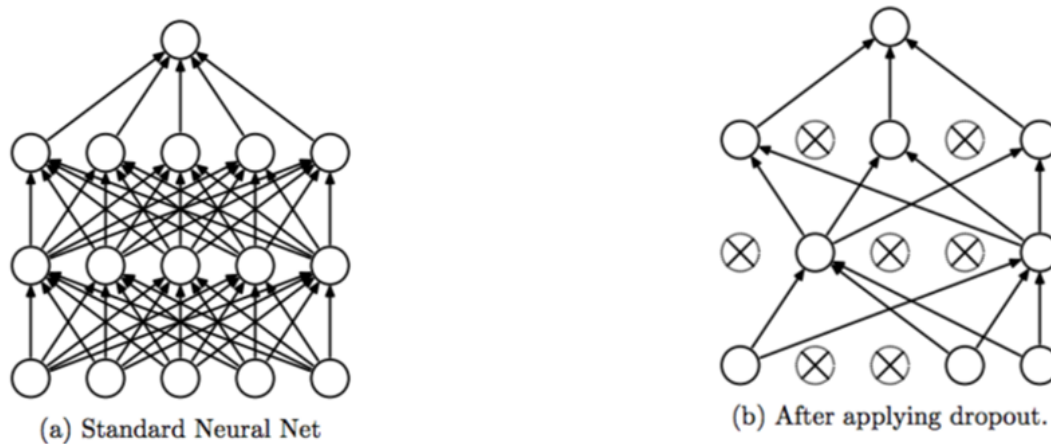


Figure 14 The Dropout Layer

[12] Dropout is a technique where randomly selected neurons are ignored during training. They are “dropped-out” randomly. This means that their contribution to the activation of downstream neurons is temporally removed on the forward pass and any weight updates are not applied to the neuron on the backward pass.

As a neural network learns, neuron weights settle into their context within the network. Weights of neurons are tuned for specific features providing some specialization. Neighboring neurons become to rely on this specialization, which if taken too far can result in a fragile model too specialized to the training data. This reliant on context for a neuron during training is referred to complex co-adaptations.

The effect is that the network becomes less sensitive to the specific weights of neurons. This in turn results in a network that is capable of better generalization and is less likely to overfit the training data.

Once the data goes through the last LSTM layer, we pass the parameters through a Dense

network with a softmax layer that predicts three different values for the FVC based on the three different quantiles.

We pass these three predicted values in our custom loss function, that, takes sigma (or std. deviation) as the difference between the upper and the lower FVC values to calculate the metric.[27]

3.8 TRAINING THE MODEL

Inspired by multiple Kaggle solutions, we decided that it makes more sense to use K-fold for our training. Why let data dependencies mess with the training of the network. Therefore, we took $K=10$, and set the model to train.

For each fold, we would first calculate the base Laplace log likelihood metric so that we can understand how well our model is working relative to the data. [How we do this mentioned in the literature review]. Once the model would be trained, the remaining data would be used for validation. Even for validation, we would first calculate the base Laplace log likelihood metric and compare our model's performance.

We were aiming the validation LLL metric to be less between the range -6.0 to -6.5.

Remember, in this model we are not using accuracy to evaluate the performance of our model, but we are using the likelihood.

```

Fold No. 1
Training....
Training score to beat: 8.027305603027344
43/43 [=====] - 1s 4ms/step - loss: 48.7594 - score: 6.6686
[48.75938415527344, 6.668625354766846]
Validation score to beat: 8.066678047180176
5/5 [=====] - 0s 5ms/step - loss: 50.1982 - score: 6.6983
Fold No. 2
Training....
Training score to beat: 8.03429126739502
43/43 [=====] - 1s 4ms/step - loss: 215.6907 - score: 8.1430
[215.69065856933594, 8.142995834350586]
Validation score to beat: 8.006591796875
5/5 [=====] - 0s 5ms/step - loss: 209.2196 - score: 8.1283
Fold No. 3
Training....
Training score to beat: 8.031084060668945
43/43 [=====] - 1s 4ms/step - loss: 214.6587 - score: 8.1415
[214.65867614746094, 8.141528129577637]
Validation score to beat: 8.038658142089844
5/5 [=====] - 0s 5ms/step - loss: 218.4835 - score: 8.1466
Fold No. 4
Training....
Training score to beat: 8.031888008117676
43/43 [=====] - 1s 4ms/step - loss: 32.9685 - score: 6.2859
[32.96852493286133, 6.285883903503418]
Validation score to beat: 8.033307075500488
5/5 [=====] - 0s 5ms/step - loss: 37.4968 - score: 6.3777
Fold No. 5
Training....
Training score to beat: 8.03518009185791
43/43 [=====] - 1s 4ms/step - loss: 59.7585 - score: 6.8867
[59.7585334777832, 6.886720180511475]
Validation score to beat: 8.001657485961914
5/5 [=====] - 0s 4ms/step - loss: 57.2611 - score: 6.8653
Fold No. 6
Training....
Training score to beat: 8.032129287719727
43/43 [=====] - 1s 4ms/step - loss: 59.2943 - score: 6.9167
[59.29428482055664, 6.9167160987854]
Validation score to beat: 8.027667045593262
5/5 [=====] - 0s 4ms/step - loss: 55.8665 - score: 6.8709
Fold No. 7
Training....
Training score to beat: 8.02800464630127
43/43 [=====] - 1s 4ms/step - loss: 213.9129 - score: 8.1394
[213.91294860839844, 8.139429092407227]
Validation score to beat: 8.067682266235352
5/5 [=====] - 0s 5ms/step - loss: 225.2055 - score: 8.1589
Fold No. 8
Training....
Training score to beat: 8.032888412475586
43/43 [=====] - 1s 7ms/step - loss: 48.6226 - score: 6.6752
[48.62260818481445, 6.675217628479004]
Validation score to beat: 8.025102615356445
5/5 [=====] - 0s 7ms/step - loss: 54.8337 - score: 6.7604
Fold No. 9
Training....
Training score to beat: 8.034184455871582
43/43 [=====] - 1s 4ms/step - loss: 45.9746 - score: 6.6225
[45.974586486816406, 6.622529983520508]
Validation score to beat: 8.004183769226074
5/5 [=====] - 0s 5ms/step - loss: 50.1161 - score: 6.6603
Fold No. 10

```

Figure 15 Model Training

The above figure is the screenshot of our model training for 10-folds on 1200 epochs each. Let us take a deeper look how did we evaluate the best fold.

In the first fold, the training score to beat is 8.027 and our training score is 6.66. This means that our model is performing well, since our training score is lesser than 8.027. Fold 1 even performed well on validation, as the validation score to beat was 8.066 and our model has a score of 6.69. This means that the model is performing excellently even on the data it has never seen.

But when we look at fold-2 the training score to beat is 8.03 and our training score is 8.14. This means that our model's predictions are obsolete and its predictions are even worse than average of the FVC values.

To better look at and understand the performance of each fold, we can look at the figure.

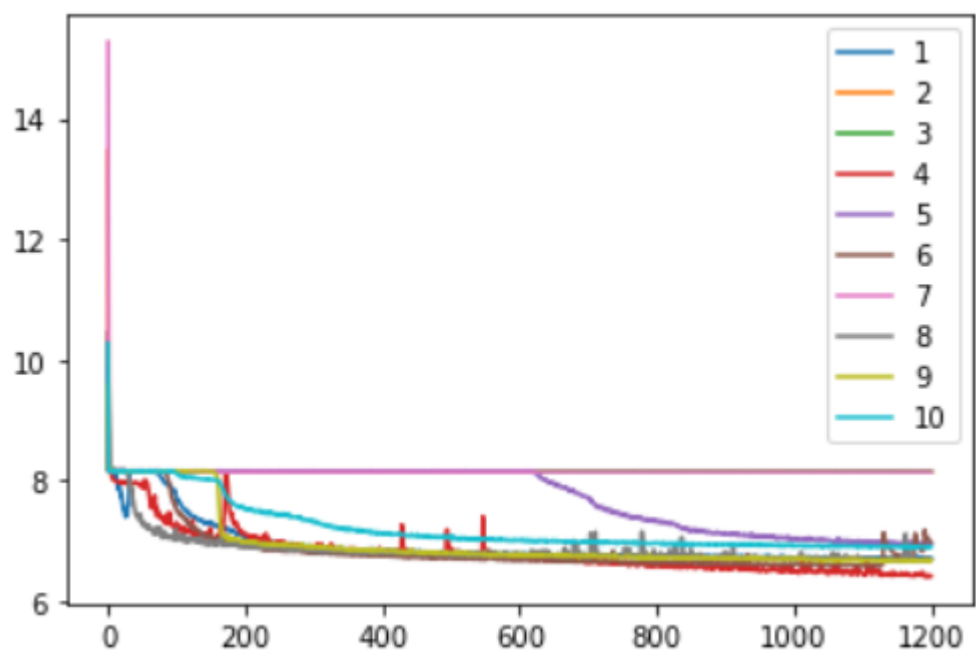


Figure 16 Training score progression of each fold

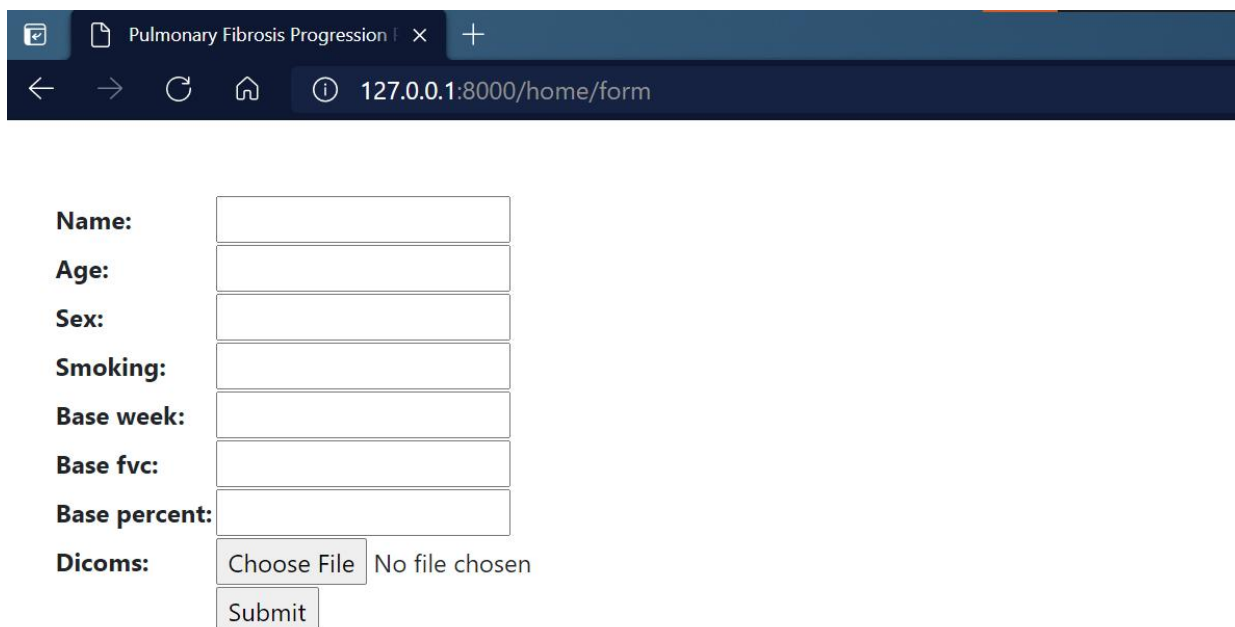
Looking at the figure it can be concluded that fold 4 did the best, as it has the lowest score. Remember, the more the predictions are tending over zero, the better it is.

4. IMPLEMENTATION

4.1 CAD SYSTEM DEVELOPMENT

To deploy our machine learning model, we built a Django portal. Doctors will be able to enter the patient's details along with the patient's DICOM files and the application will give the predict the FVC values for the next 50 weeks. We are also keeping in mind that the farther the week, the less accurate the predictions will be, therefore, we are limiting the predictions to just 50 weeks only.

Here are the screenshots of the application:



The screenshot shows a web browser window with the title "Pulmonary Fibrosis Progression | ×". The address bar displays "127.0.0.1:8000/home/form". The form contains the following fields and controls:

- Name:**
- Age:**
- Sex:**
- Smoking:**
- Base week:**
- Base fvc:**
- Base percent:**
- Dicoms:** No file chosen
-

Figure 17 The doctor will fill the above form and upload the CT scan of the patient.

Once the doctors fill the submit button, the following processes will be executed.

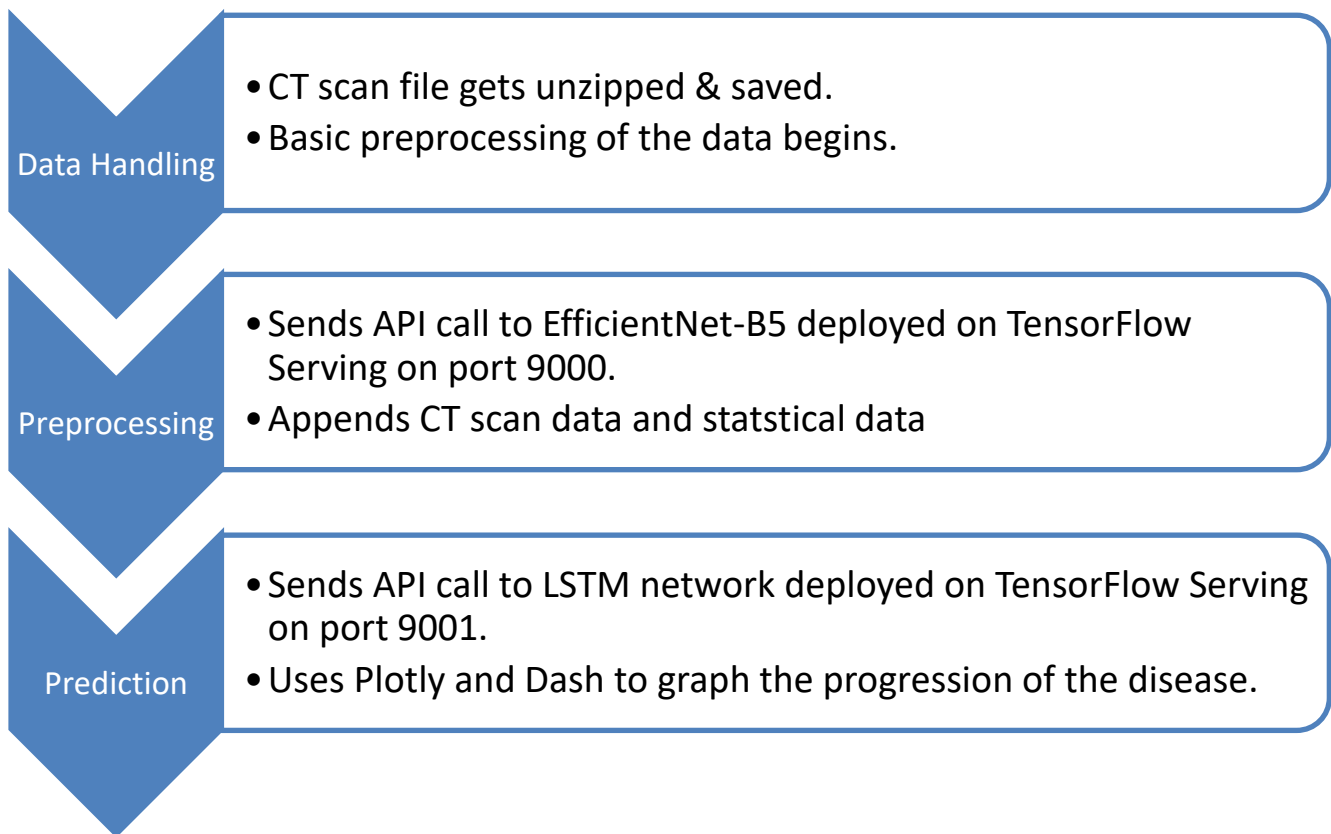


Figure 18 The process flow of the web application



Figure 19 Once the CT scan files gets uploaded, the doctor will be able to see the prediction graph.

5. RESULTS AND DECLARATION

Doctors have several ways to treat pulmonary fibrosis, including the use of medications, oxygen therapy, non-medical treatments (such as exercise), and even lung transplantation. Doctors usually recommend CT scans for finding out lung patterns and checking out FVC scores of the patients.

The Feature Extraction from the DICOM files we did was using the EfficientNet. The whole project uses a neural network for creating the model, we used simple neural network architecture and the LSTM architecture for training the model. The simple neural network model had less validation accuracy and required execution for a greater number of epochs whereas the LSTM model gave very close validation and training accuracy. Hence, the final model selected was LSTM.

5.1 CONCLUSION & FUTURE WORK

Now that both the aims of the model have been achieved, we can compare our validation accuracy to that of other solution that we discussed in the literature review.

EfficientNet is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient.

In general, the **EfficientNet** models achieve both **higher** accuracy and **better** efficiency over existing CNNs, reducing parameter size and FLOPS by an order of magnitude. ... In particular, our **EfficientNet-B7** achieves new state-of-the-art 84.4% top-1 / 97.1% top-5 accuracy, while being 8.4x smaller than the best existing CNN.

This project can be significantly used in multiple ways in the real world, especially in Hospitals and other Centre for tracking down and diagnosing with these diseases. One such use case can be the Hospital Management System where the Doctors and Medical Staff are trained and can diagnose the issues of the patient, in such a setting this model can help detect the disease of patients on call and give real-time diagnosis on the efficiency of the Doctors and Staff executive.

The future endeavors may include delivering customized software's for different organizations such as Hospital Chains, Diagnosis Centre and others. One of the greatest achievements is the high accuracy that the model is providing and the lifesaving solution for this disease.

Our best validation accuracy: 6.377

Solution	Best validation score
Art's first place solution	6.69
Abhishek Bhat's QuantileReg + Linear Decay solution	6.837
Boxiang Yun's solution using XGBoost	6.93
Yasufumi Nakama's LGB baseline approach	6.96

Thus, we can conclude that we have provided the best solution to the problem statement.

REFERENCES

1. <https://www.kaggle.com/artkulak/inference-45-55-600-epochs-tuned-effnet-b5-30-ep>
2. <https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/discussion/165727>
3. <https://www.kaggle.com/hfutybx/osic-feature-extract-from-ct>
4. Benjamin Le Cook, Williard G. Manning, “Thinking beyond the mean: a practical guide for using quantile regression methods for health services research,” in US National library of Medicine, 2013.
5. Nancy L Wilczynski, Douglas Morgan, R Brian Haynes, the Hedges Team-1, “An overview of the design and methods for retrieving high-quality studies for clinical care,” in US National library of Medicine, 2005.
6. F.A. Gers, J. Schmidhuber, F. Cummins, “Learning to forget: continual prediction with LSTM,” in IET: Digital Library , 1999.
7. <https://www.kaggle.com/yasufuminakama/osic-lgb-baseline>
8. Harvard Chen, “[DICOM Processing and Segmentation in Python.](#)” in Radiology Data Quest, 2017.
9. Alexander Etz, “Introduction to the Concept of Likelihood and Its Applications,” in Association of Psychological Science, 2018.
10. Nitish Srivastava and Geoffrey Hinton and Alex Krizhevsky and Ilya Sutskever and Ruslan Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” in Journal of Machine Learning Research, 2014.
11. Sakshi Indolia, Anil Kumar Goswami, S.P. Mishra, Pooja Asopa, “ Conceptual understanding of Convolutional Neural Networks – A deep learning Approach” in Procedia Computer Science, 2018.

12. Simon L F Walsh, Stephen M Humphries, Athol U Wells, Kevin K Brown, "Imaging research in fibrotic lung disease; applying deep learning to unsolved problems", in *Lancet Respir Med*, 2020.
13. He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition. CVPR,"
14. Walsh SLF, Calandriello L, Silva M, Sverzellati N., "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study," in *Lancet Respir Med* 2018.
15. Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
16. M. Anthimopoulos, S. Christodoulidis, A. Christe and S. Mougiakakou, " Classification of Interstitial Lung Disease Patterns Using Local DCT Features and Random Forest."
17. Q. Li et al., "Lung image patch classification with automatic feature learning," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* 2013.
18. K.T. Vo et al., "Multiple kernel learning for classification of diffuse lung disease using HRCT lung images," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* 2010
19. M. Gangeh et al., "A texton-based approach for the classification of lung parenchyma in ct images", *Med Image Comput Comput Assist Interv. Vol. 13(Pt 3)*
20. Blackwell, Timothy S., et al. "Future directions in idiopathic pulmonary fibrosis research. An NHLBI workshop report." *American journal of respiratory and critical care medicine* 189.2 (2014): 214-222.
21. Chua, Felix, Jack Gauldie, and Geoffrey J. Laurent. "Pulmonary fibrosis: searching for model answers." *American journal of respiratory cell and molecular biology* 33.1 (2005)
22. Czaplinski, A., A. A. Yen, and Stanley H. Appel. "Forced vital capacity (FVC) as an indicator of survival and disease progression in an ALS clinic population." *Journal of Neurology, Neurosurgery & Psychiatry* 77.3 (2006): 390-392.
23. Zappala, C. J., et al. "Marginal decline in forced vital capacity is associated with a poor outcome in idiopathic pulmonary fibrosis." *European Respiratory Journal* 35.4 (2010):

830-836.

24. Koenker, Roger, and Kevin F. Hallock. "Quantile regression." *Journal of economic perspectives* 15.4 (2001): 143-156.
25. Hao, Lingxin, Daniel Q. Naiman, and Daniel Q. Naiman. *Quantile regression*. No. 149. Sage, 2007.
26. Bottai, Matteo, Nicola Orsini, and Marco Geraci. "A gradient search maximization algorithm for the asymmetric Laplace likelihood." *Journal of Statistical Computation and Simulation* 85.10 (2015): 1919-1925.
27. Baldi, Pierre, and Peter J. Sadowski. "Understanding dropout." *Advances in neural information processing systems* 26 (2013): 2814-2822.

MEETING AGENDA-WEEK ONE (7th-12th December)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	11/12/2020	Time:	4:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective

Brainstorming different approaches for the model with pros and cons and which best satisfies our requirement.

2. Attendees

Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA-WEEK TWO (14th-19th December)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	16/12/2020	Time:	4:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective

1. Presenting the final approach and justifying our choice.
2. Seeking guidance regarding implementation

2. Attendees

Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA-WEEK THREE(18th- 23rd January)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	19/01/2021	Time:	4:30 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective

1. Seeking guidance on how to handle data larger than the RAM capacity.
2. Discussion on the how to segment the lung tissue better and faster.

2. Attendees

Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA-WEEK FOUR (25th-30th January)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	26/01/2021	Time:	4:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective

Project update meeting, discussion on the deliverables of M1.

2. Attendees

Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA-WEEK FIVE(1st -6th February)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	05/02/2021	Time:	4:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective	
1.	M1 presentation with Artika ma'am
2.	Discussion on difficulties in embedding EfficientNets in the application.

2. Attendees			
Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA- WEEK SIX(8th – 13th February)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	10/02/2021	Time:	4:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective	
Discussion on bypassing the use of EfficientNets for now.	

2. Attendees			
Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA-WEEK SEVEN (15th-20th February)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	18/02/2021	Time:	5:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	External Faculty:	Assistant Professor Artika Singh

1. Meeting Objective	
1.	Discussion on deliverables of M2
2.	Discussion on the structure of the LSTM model.
3.	Doubts with Laplace Log Likelihood

2. Attendees			
Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA- WEEK EIGHT (1st –6th March)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	04/03/2021	Time:	4:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective	
1. Project update meeting, we present a LSTM model without image data	

2. Attendees			
Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692

MEETING AGENDA- WEEK NINE (8th – 13th March)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	10/03/2021	Time:	4:00 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	Platform:	Online Meeting conducted on teams

1. Meeting Objective	
1. Discussion on how to build production ready application. 2. How to deploy big ML models to production?	

2. Attendees			
Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725

Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692
---------------	-------------------	------------------------------	------------

MEETING AGENDA- WEEK TEN (22nd – 27th March)

Meeting/Project Name:	CAD system for Lung disease		
Date of Meeting:	25/03/2021	Time:	4:30 pm
Meeting Facilitator:	Assistant Professor Supriya Agrawal	External Faculty:	Assistant Professor Artika Singh

1. Meeting Objective

To present the deployed working model and CAD software.

2. Attendees

Name	Department/Division	E-mail	Phone
Harsh Ajmera	MBA(tech)CS-Div G	Harsh.ajmera008@nmims.edu.in	9511290293
Naman Bansal	MBA(tech)CS-Div G	Naman.bansal14@nmims.edu.in	7588860475
Abhishek Jain	MBA(tech)CS-Div G	Abhishek.jain39@nmims.edu.in	8291608725
Aditya Mittal	MBA(tech)CS-Div H	Aditya.mittal61@nmims.edu.in	9079349692