PREDICTING PULMONARY FIBROSIS PROGRESSION USING DEEP LEARNING

CSE/EEE499A

Faculty:

DR. TANZILUR RAHMAN
Assistant Professor, North South University

Section:18
Group: 1

Members:

Shazzad Hasan 1530604043
Md. Tanzim Hossain 1620776042
Md. Saidur Rahman 1621529042

Submission Date: 12/08/2020

Table of Contents

Abstract:

Idiopathic pulmonary fibrosis (IPF) is a fatal lung disease characterized by an unpredictable progressive decline in lung function. Computed tomography (HRCT) has been used for the diagnosis of IPF, but not generally for monitoring purpose. The objective of our work is to develop a predictive model for the radiological progression pattern using baseline CT scans and clinical information of the patients. We will use our dataset to train and build the model for predicting IPF progression. For this work we will use convolution neural network to extract complex feature from the CT scan and for numerical and categorical data we will use multilayer perceptron in order to learn the correlation between the data. Then using fully connected layer, activation function and softmax, we will predict the progression of IPF. After that we will test our model with unseen test data in order to evaluate our model.

Motivation:

Imagine one day, your breathing became consistently labored and shallow. Months later you were finally diagnosed with pulmonary fibrosis, a progressive disease that naturally gets worse over time with no known cause and no known cure, created by scarring of the lungs. If that happened to you, you would want to know your prognosis. That's where a troubling disease becomes frightening for the patient. Outcomes can range from long-term stability to rapid deterioration, Natural history of IPF is unknown and the prediction of disease progression at the time of diagnosis is notoriously difficult and doctors aren't easily able to tell where an individual may fall on that spectrum. Data science, may be able to aid in this prediction. If successful, patients and their families would better understand their prognosis when they are first diagnosed with this incurable lung disease. Improved severity detection would also positively impact treatment trial design and accelerate the clinical development of novel treatments.

Aim and Objective:

Lung function is assessed based on output from a spirometer, which measures the forced vital capacity (FVC), i.e. the volume of air exhaled. Our aim is to predict a patient's severity of decline in lung function based on a CT scan of their lungs, metadata, and baseline FVC as input. We want to predict the final three FVC measurements for each patient, as well as a confidence value in our prediction.

- **Patient_Week** - a unique Id formed by concatenating the Patient and Weeks columns (i.e. ABC_22 is a prediction for patient ABC at week 22)
- **FVC** - the predicted FVC in ml
- **Confidence** - a confidence value of your prediction (also has units of ml)

Background Study:

A paper titled "Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization - Random forest approach" [1] by Yu Shi was published in 2019 which is very recent. Their work is the first approach to show that it is possible to use only baseline HRCT scans to predict progression of idiopathic pulmonary fibrosis using artificial intelligence. In their paper they try to develop a novel predictive model for the radiological progression pattern of idiopathic pulmonary fibrosis using only baseline HRCT scans. First they implemented a study design and having an expert radiologist contour region of interests (ROI) at baseline scans, depending on its progression status in follow-up visits. Then they integrated the feature selection with prediction by developing an algorithm using a wrapper method that combines quantum particle swarm optimization to select a small number of features with random forest to classify early patterns of progression. They compare their result with other popular wrappers and non-wrapper methods, i.e. smoothly clipped absolute deviation (SCAD), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), neural network (NNET). Their proposed model yields an overall accuracy rate of 82.1% which is superior to other feature selections and classification methods mentioned above.

Next paper titled "Idiopathic Pulmonary Fibrosis: Gender-Age-Physiology Index Stage for Predicting Future Lung Function Decline" [3] by Margaret L. Salisbury. In the paper they showed that patients with idiopathic pulmonary fibrosis (N = 657) were identified retrospectively at three tertiary referral centers, and baseline GAP stages were assessed. Mixed models were used to describe average trajectories of FVC and diffusing capacity of the lung for carbon monoxide (DLCO). Multivariable Cox proportional hazards models were used to assess whether declines in pulmonary function $\geq$ 10% in 6 months predict mortality after accounting for GAP stage. They found that over a 2-year period, GAP stage was not associated with differences in yearly lung function decline. After accounting for stage, a 10% decrease in FVC or DLCO over 6 months independently predicted death or transplantation. Patients with GAP stage 2 with declining pulmonary function experienced a survival profile similar to patients with GAP stage 3, with 1-year event-free survival of 59.3%. They came to a conclusion that baseline GAP stage predicted death or lung transplantation but not the rate of future pulmonary function decline.

Dataset:

We collected our dataset from Open Source Imaging Consortium (OSIC) which is a non-profit, co-operative effort between academia, industry and philanthropy. The dataset contains a baseline chest CT scan and associated clinical information for a set of patients. A patient has an image acquired at time Week = 0 and has numerous follow up visits over the course of approximately 1-2 years. The first problem we face with the data is that the relative timing of FVC measurements varies widely. The timing of the initial measurement relative to the CT

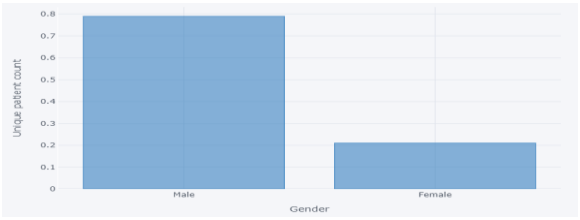
scan and the duration to the forecasted time points differ for each patient. Features of the dataset are:

- **Patient** - a unique Id for each patient (also the name of the patient's DICOM folder)
- **Weeks** - the relative number of weeks pre/post the baseline CT (may be negative)
- **FVC** - the recorded lung capacity in ml
- **Percent** - a computed field which approximates the patient's FVC as a percent of the typical FVC for a person of similar characteristics
- **Age** - Patient's age
- **Sex** - Patient's gender
- **Smoking Status**

All the related information about metadata and data visualization are given below:

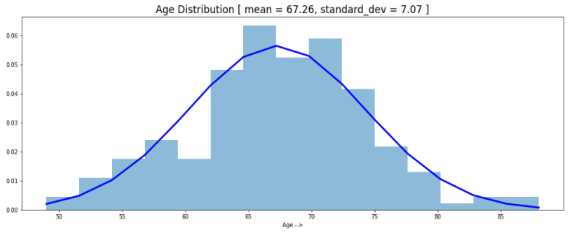
- Total 1549 patients
- 176 patients are unique
- 33,026 files/images, 176 folders/patients
- 187.0 average files/images per patient
- 1,018 max files/images per patient

Gender Distribution



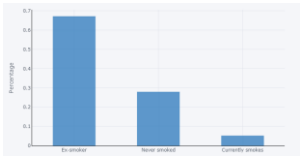
Almost 78% are Male and 22% are Female

Age Distribution



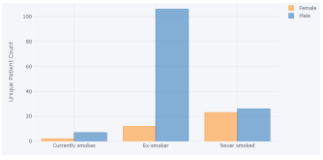
Age distribution varies from 49 years to 88 years.

Smoking Status Distribution



67% of the patients are ex-smokers, 28% never smoked and 5% are still smoking

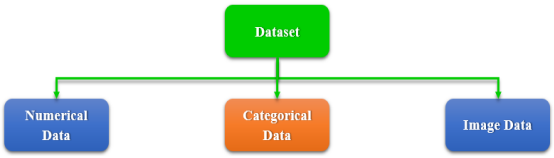
Gender vs Smoking Status



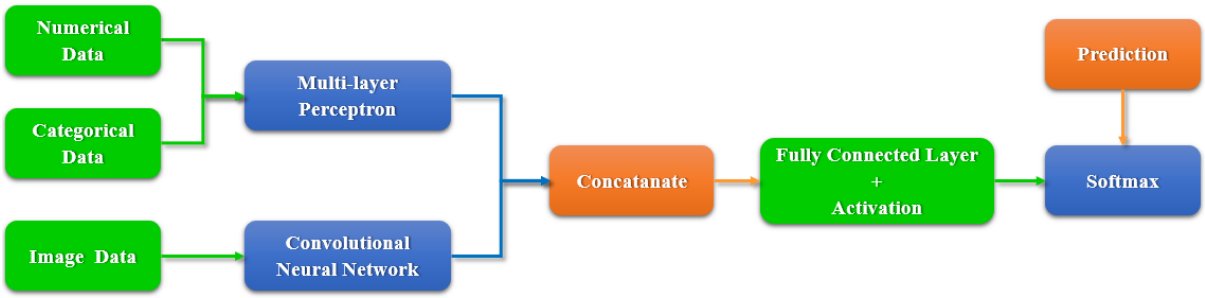
Age Distribution vs Smoking Status

The dataset contains mixed data.

- Numeric: Patient, Weeks, FVC, Age
- Categorical: Sex, Smoking status
- Image: CT scan



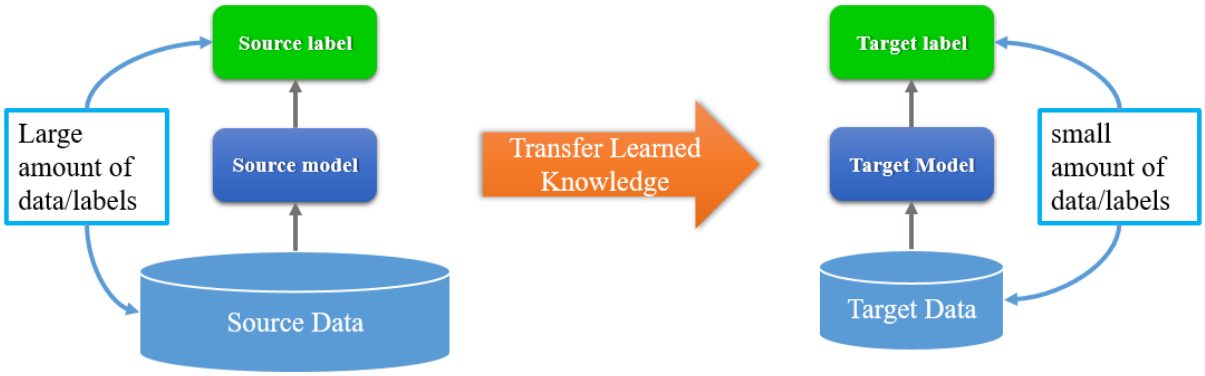
METHODOLOGY:



Approach

In this project our main objective is to predict the progression of idiopathic pulmonary fibrosis. But the prediction is notoriously difficult. So, we proposed an approach to build a model that can hopefully predict the progression of idiopathic pulmonary fibrosis.

For our work we will use transfer learning. In practice, very few people train an entire convolutional neural network from scratch, because it is relatively rare to have a dataset of sufficient size. Instead, it is common to pretrain a CNN on a very large dataset and then use the CNN either as an initialization or a fixed feature extractor for the task of interest.



Transfer Learning

The basic premise of transfer learning is simple: take a model trained on a large dataset and transfer its knowledge to a smaller dataset. For working with a CNN, we freeze the early convolutional layers of the network and only train the last few layers which make a prediction. The idea is the convolutional layers extract general, low-level features that are applicable across images — such as edges, patterns, gradients — and the later layers identify specific features within an image.

Image Preprocessing is the most important step of working with image data. In our case the image data is the CT scan. During image preprocessing, we simultaneously prepare the images for our network and apply data augmentation to the training set.

With our data in shape, we next turn our attention to the model. For this, we'll use a pre-trained convolutional neural network. There are number of models that have already been trained on millions of images from 1000 classes in Imagenet (e.g. Alexnet, VGG, ResNet, SqueezeNet, DenseNet etc.).

We will use multilayer perceptron in order to process numerical and categorical data. Multilayer perceptrons train on a set of input-output pairs and learn to model the correlation between those inputs and outputs. In order to minimize error multilayer perceptron training involves adjusting the parameters, or the weights and biases of the model.

After that we will concatenate them and process them through fully connected layer. Then we will use softmax to predict the progression in the form of FVC value and we will find confidence value for every predicted FVC value. Softmax function, an activation function that turns numbers also known as logits into probabilities that sum to one. Softmax function outputs a vector that represents the probability distributions of a list of potential outcomes. Softmax turn logits into probabilities by taking the exponents of each output and then normalize each number by the sum of those exponents so the entire output vector adds up to one. It computes softmax cross entropy between logits and labels.

Software & Tools:

The list of software and tools that we will be using throughout this project is given below

- **Python**
  - Overall Scripting
- **PyTorch**
  - Multi-layer perceptron, ConvNet, Backpropagation, Optimization and others
- **Scikit-learn**
- **Numpy, Pandas, Matplotlib**
  - Data Visualization
- **Cuda**
  - Execution

Conclusion:

Our main objective of this project is to build a predictive model that can achieve a high accuracy rate in predicting progression of IPF. There is not much work have been done in predicting progression of IPF. So, if we can build a model with high accuracy then it will be a great use for the doctor, clinicians and most importantly for the patient.

References:

[1] Shi, Y., Wong, W.K., Goldin, J.G., Brown, M.S. and Kim, G.H.J., 2019. Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization-Random forest approach. Artificial intelligence in medicine, 100, p.101709.

[2] Kim, G.H.J., Weigt, S.S., Belperio, J.A., Brown, M.S., Shi, Y., Lai, J.H. and Goldin, J.G., 2020. Prediction of idiopathic pulmonary fibrosis progression using early quantitative changes on CT imaging for a short term of clinical 18–24-month follow-ups. European Radiology, 30(2), pp.726-734.

[3] Salisbury, M.L., Xia, M., Zhou, Y., Murray, S., Tayob, N., Brown, K.K., Wells, A.U., Schmidt, S.L., Martinez, F.J. and Flaherty, K.R., 2016. Idiopathic pulmonary fibrosis: gender-age-physiology index stage for predicting future lung function decline. Chest, 149(2), pp.491-498.