Analysis of Idiopathic Pulmonary Fibrosis through Machine Learning Techniques

Upasana Chutia*, Anand Shanker Tewari†, and Jyoti Prakash Singh‡
Department of Computer Science and Engineering, National Institute of Technology Patna, India
Email: upasanac.phd19.cs@nitp.ac.in*, anand@nitp.ac.in†, jps@nitp.ac.in‡

2021 8th International Conference on Smart Computing and Communications (ICSCC) | 978-1-7281-9687-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSCC51209.2021.9528243

***Abstract*—Few diseases are hard to detect and life-threatening as well, and Pulmonary Fibrosis (PF) is one of them. PF is a chronic disorder that leads to progressive scarring of the lungs, and we can say that PF is Idiopathic Pulmonary Fibrosis (IPF) because the cause of the disease is unknown. 50,000 fresh cases per year are diagnosed with PF, which is likely to increase. With machine learning and deep learning, we can predict the lung function decline of a patient suffering from IPF. This prediction will improve the medication process and will increase the longevity of the patient. Early detection of IPF is crucial as it increases the morbidity and mortality rate and healthcare costs. We have predicted IPF in the early stages using forced vital capacity (FVC) records of different patients. FVC is the amount of air that we can exhale from our lungs after taking a deep breath. We have created a Multiple-Quantile Regression model to detect a decline in lung function using CNN. With this approach, the cross-validation accuracy of prediction is 92 per-cent.**

***Index Terms*—FVC, Idiopathic Pulmonary Fibrosis(IPF), Machine learning, Deep Learning, Multiple Quantile Regression, Elastic net, CNN**

I. Introduction

Pulmonary fibrosis is a relatively uncommon condition. Unlike diabetes, heart disease, cancer, or any other well-known disease that are somewhat detectable and curable to a certain extent, most people are not aware of pulmonary fibrosis(PF). People only understand the disease when they or their close ones are diagnosed with this disease. There are multiple words that we have used to describe patients with pulmonary fibrosis, Usual Intestinal Pneumonia(UIP), Diffuse Parenchymal Lung Disease, Nonspecific Interstitial Pneumonia(NSIP) [7][14]. Pulmonary fibrosis is an umbrella term that encompasses over a hundred different conditions. Researchers have divided PF's cause into known, unknown, granulomatous and miscellaneous, and some known causes are environmental condition and adverse drug reactions[11]. IPF and NSIP come under the unknown cause of pulmonary fibrosis[7].

Fibrosis thickens the interstitium, because of which the air secs get stiff and makes the lungs challenging to breathe. When a patient tries to take a big deep breath, he/she cannot get in as much air as someone who has normal lungs, so the amount of oxygen that gets into the lung is lower. Some significant signs and symptoms of pulmonary fibrosis are chronic coughs, cold that will not go away, shortness of breath, fatigue, low oxygen saturation at rest or activity and clubbing of the fingers. Few major tests help us decide if someone has pulmonary fibrosis by checking the patientâs lung function.

There are two main matrices for pulmonary fibrosis test. First, Forced Vital Capacity (FVC), often reduced in patients who have pulmonary fibrosis. The second test, Decreased Diffusing Capacity for Carbon-monoxide (DLCO)[15], where the oxygen from the air is trying to get into the bloodstream, is checked. Forced Vital Capacity (FVC) has been used for a long time to assess pulmonary fibrosis functional status. The study confirms that even a minimum change in FVC causes more severe damage to the patients and increases the mortality rate[1]. A Minimal 5 per-cent to 10 per-cent changes in FVC over two years increases the mortality rate over one subsequent year.

In this paper, we have created a machine learning model that will predict the decline in the lung function of any patient diagnosed with IPF. The prediction improves the life expectancy of the patient and the whole medication process. We have created this model using quantile regression and compared it with elastic net and we are able to predict the decline in lung function with an accuracy of 92 per-cent. Our contributions here in this paper are:

- We have applied quantile regression model after analysing and pre-processing of the data.
- We have applied elastic net model for prediction.
- We have analysed the result, and we have gained 92 per-cent accuracy in cross-validation set with quantile regression.

In the next section, we have discussed the related work followed by methodology, and in section fourth, we have discussed the result. In section fifth, we have elaborated our research conclusions and future work.

II. Related work

D. Bois et al. estimated the Minimal Clinically Important Difference (MCID) in patients with IPF. It shows that a slight change in FVC, i.e. (2-6%), can cause massive damage to a patient. They studied 1156 patients where baseline FVC and other functional status measures were calculated at baseline and took 24-weeks gaps after that. They have used both anchor-based and distribution-based methods to calculate the MCID. For distribution-based methods, they used

978-1-7281-9687-9/21/$31.00 © 2021 IEEE

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 20,2025 at 05:39:55 UTC from IEEE Xplore. Restrictions apply.

Standard Error of Measurement (SEM). Anchor-based methods included the criterion-referencing and patient-referencing approaches[2].

C. Wang et al. developed three FVC prediction models for Chronic Obstructive Pulmonary Disease (COPD) based on the Support Vector Regression technique. Best model showed 95% accuracy on the test data set. A one-tailed test was used to evaluate the significant difference between the actual FVC and predicted FVC[5]. J. Chen et al. used multi-output support vector regression to predict pulmonary disease prognosis based on FVC and FEV1 with inflammatory parameters and the patients' demographic details[8]. COPD is also a chronic disease like IPF, but the survival rate of COPD patients is high compared to IPF patients[6].

S. Mandal et al. predict the FVC based on the patient's previous FVC records. They have compared three regression techniques, namely elastic net, quantile regression and ridge regression [3].

J. A. Bjoraker et al. did a survey where they showed that the median survival of patients diagnosed with IPF is only 3 to 6 years. Figure 1 shows the difference between expected life and actual life after IPF diagnosis [4]. SM moon et al. found a correlation between FVC and exercise capacity of 6 minutes walking distance (6MWD) with diseased pulmonary patients. They used logistic regression to find the association between FVC and 6MWD. They found a positive association between FVC and 6MWD, with the patients having very severe pulmonary disease[9].

Even though some medications are also available in the market, viz. Pirfenidone and Nintedanib, but they have their adverse effects[12]. Anorexia and nausea are the common ADRs associated with Nintedanib, whereas rash, anorexia gastrointestinal disturbance the ADRs associated with Pirfenidone. John Hutchinson et al. conducted a survey where they collected the data from different countries. It was found that the IPF case is more common in North America and Europe and lower in East Asia and South America. The number of patients is more likely to increase along with time[13].

Pulmonary fibrosis is one of the serious concern among the COVID-19 patient as well. SARS-CoV-2 causes pneumonia, and if pneumonia is not control within time, it leads to a fibrotic lung. Study shows that 25.5% - 60.0% recovered patients from SARS diagnosed with pulmonary fibrosis.

III. Methodology

A. Data and Pre-processing

We have collected the data from OSIC (Open Source Imaging Consortium) available in Kaggle[17]. The dataset has FVC records of patients of different weeks, the total number of records is 1554, and the number of unique patients is 176, out of which 79% are male patients and 21% are female patients. The age is between 49 to 88, and the average age is 67.2. The dataset also has the patientsâ smoking status, as smoking causes IPF [10], [16]. Where 118 patients are ex-smoker, 49 patients never smoked, and 09 patients were smokers. Figure 2 shows the gender status, and figure 3 shows the patients' smoking status in the dataset. Figure 4 shows the FVC records of different weeks of three different patients differentiate by their smoking status viz. ex-smoker, never smokes, and smokers. There are no missing values in the dataset.

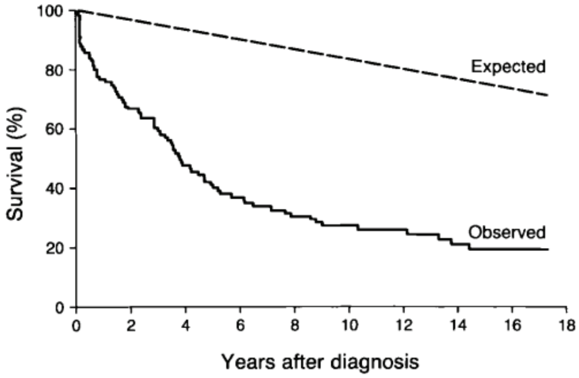


Fig. 1: FVC records of different weeks of different 3 patients[4]

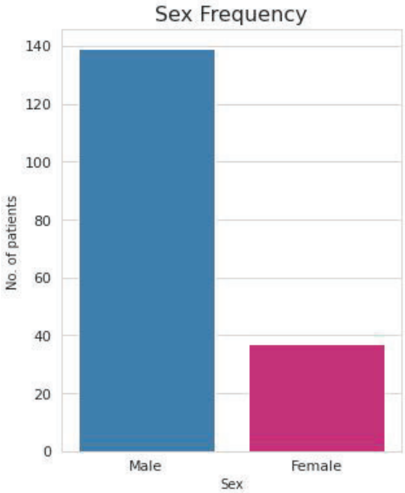


Fig. 2: Gender frequency of OSIC dataset

In the pre-processing phase we have converted nominal data to numerical data for various columns like gender and smoking status. We have merged train and test data so that we can apply K-fold cross-validation.

The dataset also contains CT scan images of each patients. The 0^{th} week of FVC record is the week when the HRCT scan images were taken from the patients. Figure 5 shows the axial, sagittal and coronal view of CT images of four different patients.

B. Quantile regression

The Quantile regression extends standard linear regression. It models conditional quantiles of dependent variables. In

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 20,2025 at 05:39:55 UTC from IEEE Xplore. Restrictions apply.

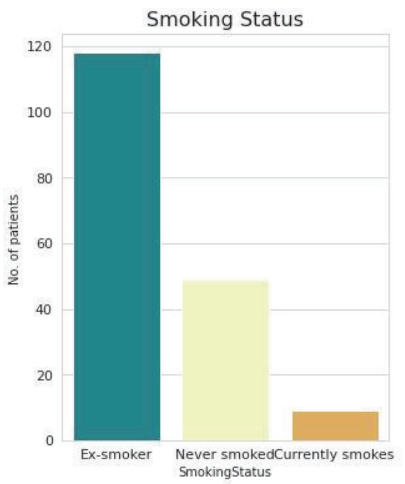


Fig. 3: Smoking status of OSIC daatset

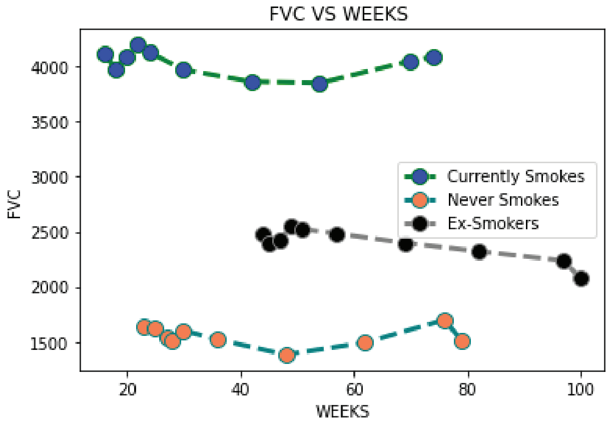


Fig. 4: FVC records of different weeks of different 3 patients

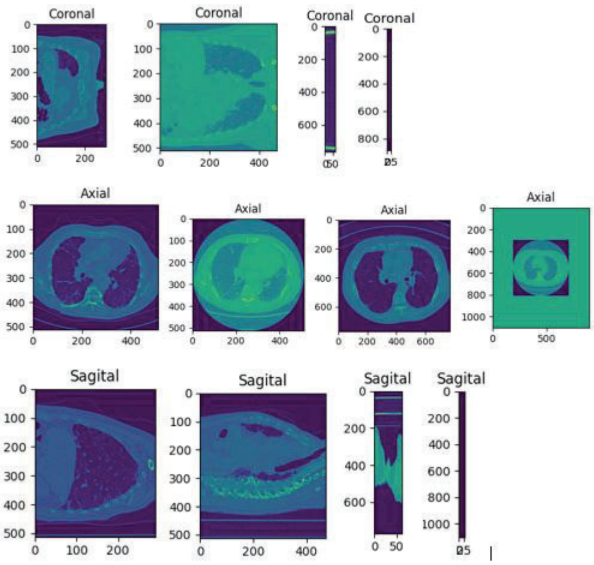


Fig. 5: The axial, sagittal and coronal view of HRCT images of four patients

contrast to Conditional Ordinary Least Square (COLS) models, the Quantile Regression Model (QRM) estimates the outcome variable's conditional median. It gives a more comprehensive picture of the effect of the independent variables on the dependent variables.

The equation of the quantile regression for the τ is illustrated in equation 1:

$$Q_{\tau}(Y_i) = \beta_0(\tau) + \beta_1(\tau)X_{i1} + ... + \beta_p(\tau)X_{ip} \tag{1}$$

Where, i = 1,..,n

The beta coefficient acts as a function rather than constants and affects the quantile. Finding the values of these betas follows a similar procedure, as regular linear quantization, but, here we have to reduce the median absolute deviation.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\rho_{\tau}(Y_i - (\beta_0(\tau) + \beta_1(\tau)X_{i1} + ... + \beta_p(\tau)X_{ip})) \tag{2}$$

Here ρ is the check function that gives asymmetric weights to the error depending on the quantile and the overall sign of the error.

C. Elastic net

Elastic net combines shrinkage from Lasso (L_1) and Ridge (L_2) regression. It combines the advantage of both the L_1 and L_2 norms. The primary advantage of using the elastic net is we don't need any assumption of the dependent variable, and it is excellent in handling data sparsity and multicollinearity.

In ridge regression or L_2 regularization, we add the sum of the squares of the parameters as a penalty term to the main loss function and give some importance of some λ to the additional term.

The objective function of the ridge regression is illustrated in equation two :

The Main Loss Function + λ Sum of Squares of the parameters

$$\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda\sum_{i=1}^{n} w_i^2 \tag{3}$$

In Lasso regression of L_1 regularization, we add the sum of absolute values of the parameters as a penalty term to the main loss function and give some weightage of some λ to the additional term.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 20,2025 at 05:39:55 UTC from IEEE Xplore. Restrictions apply.

The objective function of lasso regression is defined in equation three :

The Main Loss Function + λ Sum of Absolute Values of parameters

$$\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n} \mid w_i \mid \tag{4}$$

Elastic Net Regularization:

When some parameters are not important, it makes sense to use Lasso to get rid of those permeates that are least important in determining the target variable. The Ridge regression does not lead the sparsity, but it shrinks the parameters when there is a high correlation between them.

The elastic net is useful when we don't know which of those two we care about more. So elastic net combines them both and try to come up with a better solution. We get the L_1 and L_2 value by grid search or hyperparameter tuning.

The objective function of elastic net regression is defined in equation four :

The Main Loss Function +λ Sum of Square Values of parameters +λ Sum of Absolute Values of parameters

$$\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{n} w_i^2 + \lambda \sum_{i=1}^{n} \mid w_i \mid \tag{5}$$

IV. Result and Discussion

In our model, we have used Convolution Neural Network (CNN) for Multiple Quantile Regression (MQR) and a linear model for Elastic Net. The implementation was performed on python. The machine's architecture for the prediction is illustrated in table 1, and we have shown the CNN architecture used in table 2. We have calculated the accuracy of the models by taking the mean of cross-validation score of each fold. The accuracy of the quantile regression model is -6.13, and the prediction accuracy is 92%. In contrast, the accuracy of the elastic net is -6.46. Figure 6 shows the different quantile levels viz. 0.25%, 0.50% and 0.75%, of the predicted FVC. We have represented the summary of different quantiles in figure 7. Analysis of both the techniques, we concluded that for the prediction of pulmonary fibrosis through FVC, quantile regression gives us a better result.

TABLE I: The architecture of the machine

Hardware	Specification
Architecture	X86 with clock frequency of 34.GHz, 16-cores
L1 cache	SBK, 4 way, 32 byte block
L2 cache	SBK, 4 way, 32 byte block
Main memory	16 GB
GPU	4GB

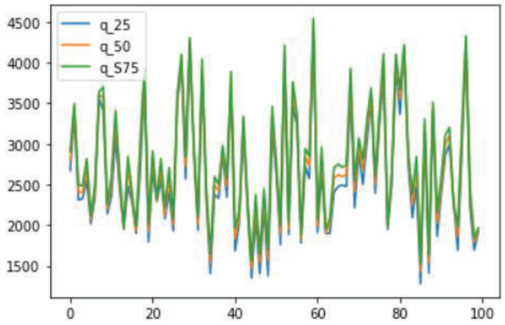
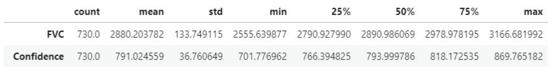


Fig. 6: Predictions for quantiles 0.25, 0.50 and 0.75



	count	mean	std	min	25%	50%	75%	max
FVC	730.0	2880.203782	133.749115	2555.639877	2790.927990	2890.986069	2978.978195	3166.681992
Confidence	730.0	791.024559	36.760649	701.776962	766.394825	793.999786	818.172535	869.765182

Fig. 7: Summary of the different quantiles

TABLE II: Architecture of CNN used in MQR model

Properties	Values
No of layers	4
No. of fold	15
Epochs	2000
Batch size	128
Learning rate	0.001
Optimizer	Adam
Total Trainable parameters	11,706
Non-trainable parameters	0

V. Conclusion

IPF is a deadly disease, and the mortality rate is high. Medical science, with the help of machine learning, may increase the longevity of the patient. We have compared two techniques for prediction, and the first one is quantile regression which uses CNN, and the other one is elastic net. We found quantile regression is giving better accuracy in cross-validation sets than the elastic net. With this prediction, we can detect various aspects or stages of IPF, and with that, the medication and health care of the patient will be improved. We have assumed that the patient is diagnosed with IPF. We are working on a prediction in which we will detect the disease just using CT scans, and this paper is a step forward towards our future work.

Acknowledgment

We are thankful to the R & D department of the National Institute of Technology Patna and the anonymous reviewers.

References

[1] Collard HR, et al, ”Changes in clinical and physiologic variables predict survival in idiopathic pulmonary fibrosis,” Am J Respir Crit Care Med 2003;168:538â542.

[2] Du Bois RM, et al, ”Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference,” American journal of respiratory and critical care medicine. 2011 Dec 15;184(12):1382-9.

[3] Mandal S, Balas VE, Shaw RN, and Ghosh A, ”Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset,” In2020 IEEE International conference on computing, power and communication technologies (GUCON) 2020 Oct 2 (pp. 861-865). IEEE.

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 20,2025 at 05:39:55 UTC from IEEE Xplore. Restrictions apply.

[4] Bjoraker JA, et al, ”Prognostic significance of histopathologic subsets in idiopathic pulmonary fibrosis,” Am J Respir Crit Care Med 1998;157: 199â203

[5] Wang C, et al, ”Predicting forced vital capacity (FVC) using support vector regression (SVR),” Physiological measurement. 2019 Feb 28;40(2):025010.

[6] [Website]https://www.healthline.com/health/managing-idiopathic-pulmonary-fibrosis/ipf-vs-copd

[7] Christe A, et al, ”Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images,” Investigative radiology. 2019 Oct;54(10):627.

[8] J. Chen, Z. Yang, Q. Yuan, D.-x. Xiong, and L.-q. Guo, ”Prediction models for pulmonary function during acute exacerbation of chronic obstructive pulmonary disease,” Physiological Measurement 41 (12) (2020) 125010.

[9] Moon SM, et al, ”Clinical impact of forced vital capacity on exercise performance in patients with chronic obstructive pulmonary disease,” Journal of Thoracic Disease. 2021 Feb;13(2):837.

[10] Maremanda KP, Sundar IK, Li D, and Rahman I. ”Age-Dependent assessment of genes involved in cellular senescence, telomere, and mitochondrial pathways in human lung tissue of smokers, COPD, and IPF: Associations with SARS-CoV-2 COVID-19 ACE2-TMPRSS2-furin-DPP4 axis,” Frontiers in pharmacology. 2020 Sep 9;11:1356.

[11] Schwaiblmair M, et al, ” Drug induced interstitial lung disease,” The Open Respiratory Medicine J. 2012; 6: 63-74.

[12] Barratt SL, et al, ”South-West of England’s experience of the safety and tolerability pirfenidone and nintedanib for the treatment of Idiopathic Pulmonary Fibrosis (IPF),” Frontiers in pharmacology. 2018 Dec 17;9:1480.

[13] Hutchinson J, Fogarty A, Hubbard R, and McKeever T, ”Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review,” European Respiratory Journal. 2015 Sep 1;46(3):795-806.

[14] MacDonald SL, et al, ” Nonspecific interstitial pneumonia and usual interstitial pneumonia: comparative appearances at and diagnostic accuracy of thin-section CT,” Radiology. 2001 Dec;221(3):600-5.

[15] van der Lee I, Zanen P, Grutters JC, Snijder RJ, and van den Bosch JM. ”Diffusing capacity for nitric oxide and carbon monoxide in patients with diffuse parenchymal lung disease and pulmonary arterial hypertension,” Chest. 2006 Feb 1;129(2):378-83.

[16] Schwartz DA, et al. ”The influence of cigarette smoking on lung function in patients with idiopathic pulmonary fibrosis,” American Journal of Respiratory and Critical Care Medicine. 1991 Sep 1;144(3):504-6.

[17] [Dataset] https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/data

Authorized licensed use limited to: PES University Bengaluru. Downloaded on October 20,2025 at 05:39:55 UTC from IEEE Xplore. Restrictions apply.