Prognosing Idiopathic Pulmonary Fibrosis with Machine Learning

Arnav Kumar

April 19, 2021

1 Introduction

Idiopathic Pulmonary Fibrosis. Idiopathic Pulmonary Fibrosis (IPF) or Cryptogenic Fibrosing Alveolitis is a disease affecting the lung base and leads to lung function decline with little to no therapies available other than lung transplant (Gross and Hunninghake, 2001; Mason et al., 1999). Although it was previously believed that the disease affects only 5 out of every 100,000 individuals, the disease is now known to be much more prevalent (Coultas et al., 1994; Mason et al., 1999; Raghu et al., 2018). While age-related, with a median diagnosis age of 66, there is no known cause (King Jr et al., 2011; Raghu et al., 2018). Recently, there have been claims that IPF occurs as a result of abnormally activated alveolar epithelial cells (King Jr et al., 2011).

Patients of IPF experience a shortness of breath, and some features of the disease include diffuse pulmonary infiltrates recognizable by radiography and varying degrees of inflammation or fibrosis (Gross and Hunninghake, 2001). Affected lung areas alternate with unaffected areas in the lung (Gross and Hunninghake, 2001). Affected areas are characterized by the differences in cell age and due to a honeycomb fibrosis pattern (Gross and Hunninghake, 2001).

The outcome of Pulmonary Fibrosis can range from rapid health declination to a healthy stability, but doctors are unable to easily diagnose the severity of the disease. There exist methods to diagnose severity, but these can be complicated and are not standardized (Robbie et al., 2017). An example of such a method is a cough scale questionnaire or a shortness of breath questionnaire (King Jr et al., 2014; Robbie et al., 2017; van Manen et al., 2016). Another method of diagnosing severity is through a functionality test known as the 6 month 6 minute Walk Distance or 6MWD test, but as the name suggests, this test is not instantaneous, and still requires the effort of trained professionals (du Bois et al., 2014; Robbie et al., 2017). On the other hand, Machine learning has been used with data from different points in time to provide a prognosis by using a software tool called CALIPER that uses radiological changes to predict IPF severity (Maldonado et al., 2014). Another case of using machine learning used computed tomography (CT) scans of the lung region and obtained an accuracy of around 76.4% or 70.7%, only outperformed 66% of doctors and only classified the severity rather than providing numerical estimates (Walsh, Calandriello, et al., 2018). An accurate prognosis of the disease will put patients at more ease, and may pave the path for any treatments that will come in the future. For this reason, it is essential that a consistent and easy method for diagnosing the severity of the disease is found.

Deep Learning Methods. Machine learning is a good fit for the task at hand because doctors can let the program run given the data, and it has been used in the past to diagnose other diseases and make predictions (Wang et al., 2010). Although machine learning has been used before for this task (du Bois et al., 2014; Maldonado et al., 2014; Robbie et al., 2017), the accuracy of the models can be improved on. Furthermore, a machine learning model could make it easier to get a prognosis.

For a disease such as IPF which is a fibrosing disease within the lungs, imaging the lungs through CT scans yields in enough insight to accurately evaluate the patients prognosis (Walsh, Devaraj, et al., 2018).

Furthermore, for injuries like neck fractures, machine learning has proven to be an improvement to the prediction performance using a method of bayesian classification (Kukar et al., 1996). For diseases like cancer, machine learning has also been used to give a prognosis and modern machine learning methods have been shown to outperform more classical methods including decision trees (Cruz and Wishart, 2006). On another note, machine learning has already been used with images of leafs to determine plant diseases and their severity, showing the ability to handle and diagnose disease severity based on a CT scan input using machine learning (Mwebaze and Owomugisha, 2016).

Question. This study aims to create a model that uses one baseline CT scan, as well as the forced vital capacity (FVC) of the lungs over the time period of one to two years. The model then predicts the FVC of the lungs for the next 3 checkups, and hence predicting the rate at which the lung condition degrades. The main questions of interest are: what is the greatest accuracy a machine learning model can attain in predicting the FVC of a IPF patient on their next 3 checkups, and what method produces this accuracy?

2 Proceduce

This study employs the use of many models machine learning models. These models are coded in python (Van Rossum and Drake, 2009) with the packages Tensorflow2 (Martin Abadi et al., 2015), Scikit-learn (Pedregosa et al., 2011), and Pandas (McKinney et al., 2010). Many of these models are modified and influenced from the work of kaggle notebooks (Consortium, 2020). These models include a linear regression, simple neural network, linear regression with auto-encoder generated features, simple neural network with auto-encoder generated features, bayesian, quantile regression, and linear decay. To begin, though, exploratory data analysis is performed, and the data is preprocesed for use by the models.

Laplace Log Likelihood Metric The use of percent accuracy cannot be employed as the model is not given a categorization task, but rather a regression task. Using percent accuracy requires the model output to be discete, not continuous. For this reason, the use of the Laplace Log Likelihood (LLL) metric is employed to measure the model accuracy. The model's FVC prediciton, the true FVC, and the model's confidence are required to calculate the LLL. (Actually, confidence is a misnomer. A higher confidence score corresponds to a greater model uncertainty.)

A LLL closer to 0 represents a model which is more accurate, but the score 0 itself is unattainable for all practical purposes (due to the nature of the metric). An example of an outstanding score would be around -6.5.

The worst score a model should get is -8.023. This score is attained as a result of always guesses the mean FVC, and always has a confidence of the standard deviation of the FVCs. Any model with a LLL lower than -8.023 is useless.

The following graph shows an example of how the model's confidence affects the metric. A confidence which is too high or too low is punished with a worse score. The local minimum describes the best metric obtainable when the predicted FVC is 2800mL, and the true FVC is 2500mL.

Linear Regression (LR). The linear regreession method relies on the assumption that the FVC can be expressed as a linear combination of the input features. In specific, the linear regression assumes the formula $y = a_1x_1 + a_2x_2 + ... + a_nx_n + b$ is true to find the patient FVC (where y is the FVC, x_1 through x_n are the input features, and the coefficients a_1 through a_n and the bias b are constants found during model fitting).

The linear regression model requires the formatting of data by including `weeks_passed` and `first_FVC` features obtained from the patient's first checkup. After the data formatting, the Scikit-learn package must be used to create a linear regressor which is then trained on the training data. This model is then used to predict the FVC for the testing data, and the model accuracy is measured.

Dense Neural Network (DNN). The dense neural network is similar to the human brain. The model contains nodes connected to each other similar to the neurons in a brain. The activation of the nodes depends on the values of the parent nodes, the weights of the connections between the nodes, and the node activation function. The input layer of nodes affects the values in the first hidden layer, which affects the values of the second hidden layer, and so on, eventually affecting the output layer values.

To begin, the data was first formatted in the same way as the linear regression, then several dense neural netwroks are made, each with a different architecture. The models are then trained on the training data, and used to predict the forced vital capacity from the testing data. Then, the model with the most accurate predicitons was chosen, and the model accuracy was calculated.

Auto-encoder. The base auto-encoder which I used to modify the exisiting mehtods was created by Kaggle user Welf Crozzo (Crozzo, 2020a). The idea of an auto-encoder is to use an encoder to strip an image into its elementary aspects, and to use a decoder to turn these elementary aspects back into an image through the use of a decoder. While both the encoder and decoder must be created for training purposes, we are only interested in the tabular data created by the encoder. The tabular data created by the encoder can then be used as input data for another model such as the linear regression and simple neural network models.

The encoder is loaded from Welf Crozzo's notebook, and is used to stride over the data, adding 2000 extra features based on the patient's CT scan DICOMS. Using this new input data, a linear regression and many simple neural network models are created. All these models are then trained on the training data, and used to predict the FVC for the testing data. The best simple neural network is selected and the model accuracies are calculated for the linear regression and simple neural netwrok models.

Bayesian Partial Pooling. The bayesian method is modified from Kaggle user Carlos Souza and uses partial pooling, the idea that each patient is fitted with their own individual linear curve, but all linear curves are related by a common distribution (Souza, 2020). The slope and y-intercept of the models are distributed according to a normal distribution, and the deviance of the model from the average model helps determine the confidence. In specific, each patient has their own α_i and β_i derived from a common normal distribution. Then, the FVC is predicted for the patient using the linear model $y = \alpha_i x + \beta_i$, and the confidence is found based on the amount of data known for the patient for that time range (more data means more confidence).

To begin, features were removed from the data, and the data was reformatted into a matrix completion task. Then, the partial pooling bayesian heirachical model is created and is trained.

The testing data is then similarily converted into a matrix completion task, and the model is then used to predict the FVC for the testing data. Finally, the model accuracy is calculated.

Multiple Quantile Regression. The multiple bayesian regression method is from Kaggle user Ulrich G (G, 2020). The method uses convolutional neural networks and quantile regression to determine the model confidence. The quantile regression give the first and third quantiles of the FVC, which can be used to find a spread, and hence a measure of confidence.

The multiple quantile regression requires the initial formatting of data by creating base information similarily to the linear regression. After that, the convolutional neural network is made and trained on the tabular data. The quartile diffrerence from the ground truth is then used to calculate the model confidence. Finally, the model predicts the FVC of the patients in the testing data, and the accuracy of the model is calculated.

Linear Decay Theory. The linear decay method originates from Welf Crozzo's kaggle notebook (Crozzo, 2020b). The model assumes that the FVC of the patient decays according to the formula $FVC = a.quantile(0.75)(week - week_{test}) + FVC_{test}$, and that the confidence decays according to the formula $Confidence = Percent + a.quantile(0.75)|week - week_{test}|$. A convolutional network (CNN) is then used to predict the coefficient a. Since a CNN is used, the CT scans can be analysed in this method.

Similar to the other models, the data must first be formatted. A linear decay model is then created, and a convolutional neural network is made to predict the coefficicents of the model. The convolutional neural network is trained with the training data, and is then used to predict the FVC and confidence for the testing data. Following the prediciton, the model accuracy is calculated.

3 Results

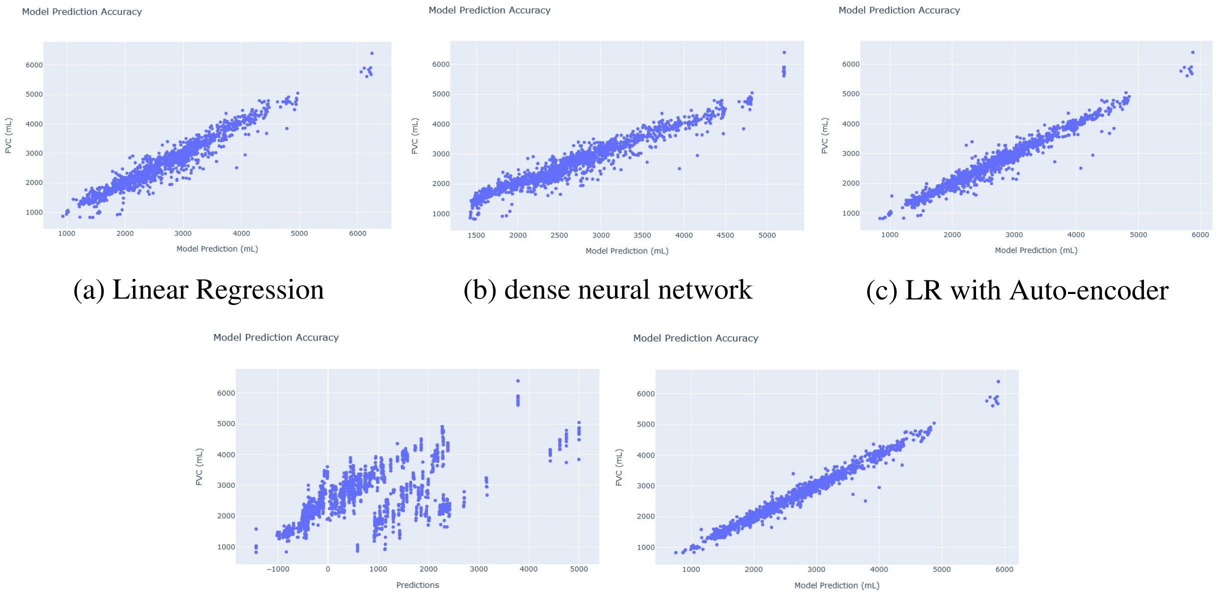
Figure 1 displays the model performance of the models analysed using the model's Laplace Log Likelihood. Training data is the same data that the model was trained on, whereas testing data is

Model	Training data	Private testing data	Public testing data
Linear Regression	-6.671	-6.867	-6.902
Dense Neural Network	-6.763	-6.888	-6.953
LR with Auto-encoder	-6.348		
DNN with Auto-encoder	-11.623		
Bayesian Partial Pooling	-6.146	-6.868	-6.909
Multiple Quantile Regression		-6.922	-6.845
Linear Decay Theory	-6.723	-6.877	-6.918

Figure 1: Laplace Log Likelihood of different models organized by dataset

data the model has never seen before. Out of the testing data, there is the public testing data, which is only around 15% of the total testing data, and there is private testing data, which consists of the other 85% of the testing data.

The two models with the auto-encoders do not have matric values for the private and public testing data due to Kaggle's time limit for submissions. Both models take a lot of time to run, and hence could not be submitted to the Kaggle competition.



(a) Linear Regression

(b) dense neural network

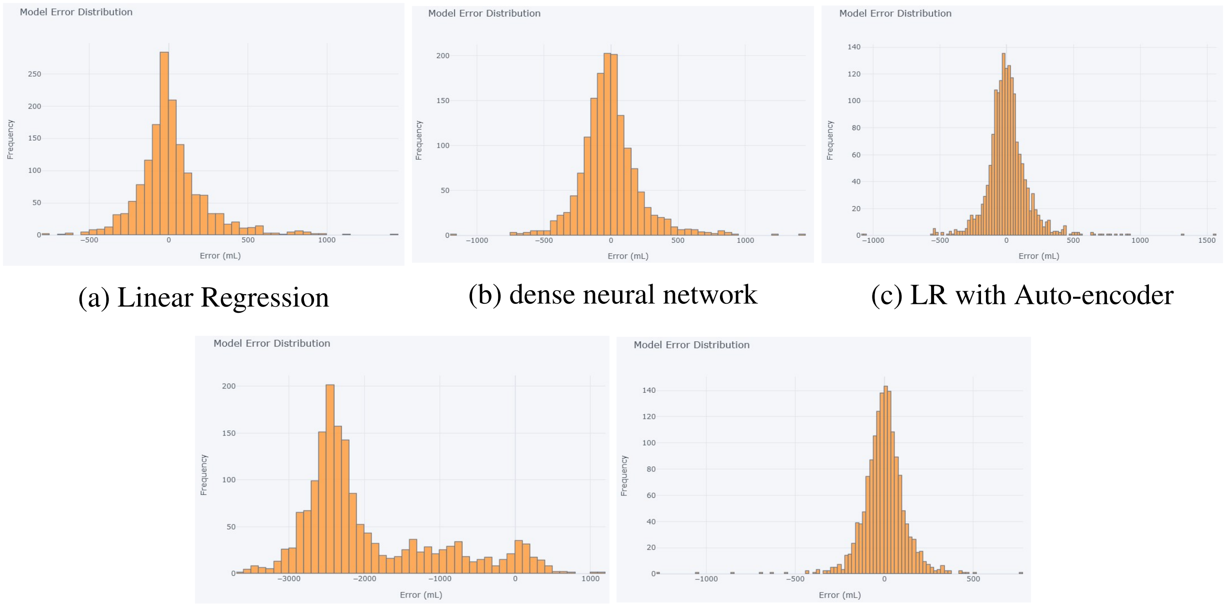
(c) LR with Auto-encoder

(d) DNN with Auto-encoder

(e) Bayesian Partial Pooling

Figure 2: Plots of True FVC vs Model Prediction

Figure 2 shows the accuracy of the predictions of several models. The true patient FVC is graphed against the model prediction, so a scattterplot closer to the line $y = x$ means the model is



(a) Linear Regression

(b) dense neural network

(c) LR with Auto-encoder

(d) DNN with Auto-encoder

(e) Bayesian Partial Pooling

Figure 3: Plots of True FVC vs Model Prediction

more accurate. In addition, figure 3 is a histogram of the errors of the models. For this reason, we desire an error which has low spread, is unimodal, and is centered at 0.

4 Conclusions

The results of the project clearly demonstrate that the DNN with Auto-encoder, DNN, and Multiple Quantile Regression models performed the worst. On the other hand, the best models were either purely or partly statistical.

5 Acknowledgements

References

Consortium, O. S. I. (2020). Osic pulmonary fibrosis progression. Kaggle. Retrieved, from https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression

Coultas, D. B., Zumwalt, R. E., Black, W. C., & Sobonya, R. E. (1994). The epidemiology of interstitial lung diseases. *American journal of respiratory and critical care medicine*, *150*(4), 967–972. https://doi.org/10.1164/ajrccm.150.4.7921471

Crozzo, W. (2020a). Image2vec: Autoencoder. https://www.kaggle.com/miklgr500/image2vec-autoencoder

Crozzo, W. (2020b). Linear decay (based on resnet cnn). https://www.kaggle.com/miklgr500/linear-decay-based-on-resnet-cnn

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*. https://doi.org/10.1177/117693510600200030

du Bois, R. M., Albera, C., Bradford, W. Z., Costabel, U., Leff, J. A., Noble, P. W., Sahn, S. A., Valeyre, D., Weycker, D., & King, T. E. (2014). 6-minute walk distance is an independent predictor of mortality in patients with idiopathic pulmonary fibrosis. *European Respiratory Journal*, *43*(5), 1421–1429. https://doi.org/10.1183/09031936.00131813

G, U. (2020). Osic-multiple-quantile-regression-starter. https://www.kaggle.com/ulrich07/osic-multiple-quantile-regression-starter

Gross, T. J., & Hunninghake, G. W. (2001). Idiopathic pulmonary fibrosis. *New England Journal of Medicine*, *345*(7), 517–525. https://doi.org/10.1056/NEJMra003200

King Jr, T. E., Bradford, W. Z., Castro-Bernardini, S., Fagan, E. A., Glaspole, I., Glassberg, M. K., Gorina, E., Hopkins, P. M., Kardatzke, D., Lancaster, L., Et al. (2014). A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis. *New England Journal of Medicine*, *370*(22), 2083–2092. https://doi.org/10.1056/NEJMoa1402582

King Jr, T. E., Pardo, A., & Selman, M. (2011). Idiopathic pulmonary fibrosis. *The Lancet*, *378*(9807), 1949–1961. https://doi.org/10.1016/S0140-6736(11)60052-4

Kukar, M., Kononenko, I., & Silvester, T. (1996). Machine learning in prognosis of the femoral neck fracture recovery. *Artificial intelligence in medicine*, *8*(5), 431–451. https://doi.org/10.1016/S0933-3657(96)00351-X

Maldonado, F., Moua, T., Rajagopalan, S., Karwoski, R. A., Raghunath, S., Decker, P. A., Hartman, T. E., Bartholmai, B. J., Robb, R. A., & Ryu, J. H. (2014). Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. *European Respiratory Journal*, *43*(1), 204–212. https://doi.org/10.1183/09031936.00071812

Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, ... Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org]. https://www.tensorflow.org/

Mason, R. J., Schwarz, M. I., Hunninghake, G. W., & Musson, R. A. (1999). Pharmacological therapy for idiopathic pulmonary fibrosis: Past, present, and future. *American Journal of Respiratory and Critical Care Medicine*, *160*(5), 1771–1777. https://doi.org/10.1164/ajrccm.160.5.9903009

McKinney, W. Et al. (2010). Data structures for statistical computing in python, In *Proceedings of the 9th python in science conference*. Austin, TX.

Mwebaze, E., & Owomugisha, G. (2016). Machine learning for plant disease incidence and severity measurements from leaf images, In *2016 15th ieee international conference on machine learning and applications (icmla)*. IEEE. https://doi.org/10.1109/ICMLA.2016.0034

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Raghu, G., Remy-Jardin, M., Myers, J. L., Richeldi, L., Ryerson, C. J., Lederer, D. J., Behr, J., Cottin, V., Danoff, S. K., Morell, F., Et al. (2018). Diagnosis of idiopathic pulmonary fibrosis. an official ats/ers/jrs/alat clinical practice guideline. *American journal of respiratory and critical care medicine*, *198*(5), e44–e68. https://doi.org/10.1164/rccm.201807-1255ST

Robbie, H., Daccord, C., Chua, F., & Devaraj, A. (2017). Evaluating disease severity in idiopathic pulmonary fibrosis. *European Respiratory Review*, *26*(145). https://doi.org/10.1183/16000617.0051-2017

Souza, C. (2020). Bayesian experiments. https://www.kaggle.com/carlossouza/bayesian-experiments

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA, CreateSpace.

van Manen, M. J., Birring, S. S., Vancheri, C., Cottin, V., Renzoni, E. A., Russell, A.-M., & Wijsenbeek, M. S. (2016). Cough in idiopathic pulmonary fibrosis. *European Respiratory Review*, *25*(141), 278–286. https://doi.org/10.1183/16000617.0090-2015

Walsh, S. L., Calandriello, L., Silva, M., & Sverzellati, N. (2018). Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: A case-cohort study. *The Lancet Respiratory Medicine*, *6*(11), 837–845. https://doi.org/10.1016/S2213-2600(18)30286-8

Walsh, S. L., Devaraj, A., Enghelmayer, J. I., Kishi, K., Silva, R. S., Patel, N., Rossman, M. D., Valenzuela, C., & Vancheri, C. (2018). Role of imaging in progressive-fibrosing interstitial lung diseases. *European Respiratory Review*, *27*(150). https://doi.org/10.1183/16000617.0073-2018

Wang, Y., Fan, Y., Bhatt, P., & Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *Neuroimage*, *50*(4), 1519–1535. https://doi.org/10.1016/j.neuroimage.2009.12.092