Predicting Idiopathic Pulmonary Fibrosis Progression

Arnav Kumar
Grade 11 Webber Academy Applied Science Student*

Correspondence to **Dr. Christian Jacob**
University of Calgary - Department of Computer Science,
Department of Biochemistry and Molecular Biology

Introduction

Idiopathic Pulmonary Fibrosis. Idiopathic Pulmonary Fibrosis (IPF) or Cryptogenic Fibrosing Alveolitis (CFA) is a disease affecting the lung base and leads to lung function decline with little to no therapies available other than lung transplant (*1, 2*). 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 (*1, 3, 4*). The disease is age-related but does not have any known cause and mainly affects older patients with the median age at diagnosis being 66 (*4, 5*). Recently, there have been claims that it is a result of abnormally activated alveolar epithelial cells (*5*). Patients 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 (*2*). Affected lung areas alternate with unaffected areas in the lung (*2*). Affected areas are characterized by the differences in cell age and due to a honeycomb fibrosis pattern (*2*).

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 meth-

*Under Supervision of Dr. Garcia-Diaz

ods to diagnose severity, but these can be complicated and are not standardized (*6*). An example of such a method is a cough scale questionnaire or a shortness of breath questionnaire (*6–8*). 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 (*6, 9*). 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 (*10*). 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 (*11*). 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 (*12*). Although machine learning has been used before for this task (*6, 9, 10*), 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 (*13*).

Furthermore, for injuries like neck fractures, machine learning has proven to be an improvement to the prediction performance using a method of bayesian classification (*14*). 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 (*15*). 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 (*16*).

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?

Procedure

This study employs the use of many models machine learning models. These models are coded in python (*17*) with the packages tensorflow (*18*), scikit learn (*19*), and pandas (*20*). Many of these models are modified and influenced from the work of kaggle notebooks (*21*). These models include a linear regression, simple neural network, linear regression with autoencoder generated features, simple neural network with autoencoder 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.

Linear Regression. 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.

Simple Neural Network. The simple 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 simple 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.

Autoencoder. The base autoencoder which I used to modify the exisiting mehtods was created by Kaggle user Welf Crozzo (*22*). The idea of an autoencoder 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 Method. 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 (*23*). 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 (*24*). 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 (*25*). 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.

References and Notes

1. R. J. Mason, M. I. Schwarz, G. W. Hunninghake, R. A. Musson, *American Journal of Respiratory and Critical Care Medicine* **160**, 1771 (1999). doi:10.1164/ajrccm.160.5.9903009.

2. T. J. Gross, G. W. Hunninghake, *New England Journal of Medicine* **345**, 517 (2001). doi:10.1056/NEJMra003200.

3. D. B. Coultas, R. E. Zumwalt, W. C. Black, R. E. Sobonya, *American journal of respiratory and critical care medicine* **150**, 967 (1994). doi:10.1164/ajrccm.150.4.7921471.

4. G. Raghu, *et al.*, *American journal of respiratory and critical care medicine* **198**, e44 (2018). doi:10.1164/rccm.201807-1255ST.

5. T. E. King Jr, A. Pardo, M. Selman, *The Lancet* **378**, 1949 (2011). doi:10.1016/S0140-6736(11)60052-4.

6. H. Robbie, C. Daccord, F. Chua, A. Devaraj, *European Respiratory Review* **26** (2017). doi:10.1183/16000617.0051-2017.

7. T. E. King Jr, *et al.*, *New England Journal of Medicine* **370**, 2083 (2014). doi:10.1056/NEJMoa1402582.

8. M. J. van Manen, *et al.*, *European Respiratory Review* **25**, 278 (2016). doi:10.1183/16000617.0090-2015.

9. R. M. du Bois, *et al.*, *European Respiratory Journal* **43**, 1421 (2014). doi:10.1183/09031936.00131813.

10. F. Maldonado, *et al.*, *European Respiratory Journal* **43**, 204 (2014). doi:10.1183/09031936.00071812.

11. S. L. Walsh, L. Calandriello, M. Silva, N. Sverzellati, *The Lancet Respiratory Medicine* **6**, 837 (2018). doi:10.1016/S2213-2600(18)30286-8.

12. Y. Wang, Y. Fan, P. Bhatt, C. Davatzikos, *Neuroimage* **50**, 1519 (2010). doi:10.1016/j.neuroimage.2009.12.092.

13. S. L. Walsh, *et al.*, *European Respiratory Review* **27** (2018). doi:10.1183/16000617.0073-2018.

14. M. Kukar, I. Kononenko, T. Silvester, *Artificial intelligence in medicine* **8**, 431 (1996). doi:10.1016/S0933-3657(96)00351-X.

15. J. A. Cruz, D. S. Wishart, *Cancer informatics* **2**, 117693510600200030 (2006). doi:10.1177/117693510600200030.

16. E. Mwebaze, G. Owomugisha, *2016 15th IEEE international conference on machine learning and applications (ICMLA)* (IEEE, 2016), pp. 158–163. doi:10.1109/ICMLA.2016.0034.

17. G. Van Rossum, F. L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

18. M. Abadi, *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.

19. F. Pedregosa, *et al.*, *Journal of Machine Learning Research* **12**, 2825 (2011).

20. W. McKinney, *et al.*, *Proceedings of the 9th Python in Science Conference* (Austin, TX, 2010), vol. 445, pp. 51–56.

21. O. S. I. Consortium, Osic pulmonary fibrosis progression (2020).

22. W. Crozzo, Image2vec: Autoencoder (2020).

23. C. Souza, Bayesian experiments (2020).

24. U. G, Osic-multiple-quantile-regression-starter (2020).

25. W. Crozzo, Linear decay (based on resnet cnn) (2020).