Predicting Idiopathic Pulmonary Fibrosis Progression Research Proposal

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

How difficult is it to give an accurate prognosis of Idiopathic Pulmonary Fibrosis? This study implements a Machine Learning Model to accurately estimate how quickly a case of the disease deteriorates. Using CT scans of the lungs from previous doctor checkups along with metadata, the model predicts the conditions of the lungs on the next three checkups. The model was made using an ensemble of various deep learning, and statistical learning methods to attain a high accuracy.

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

*Under Supervision of Dr. Garcia-Diaz

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 methods 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 know 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 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 Computed Tomography 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 lie 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

The model will use 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. What is the greatest accuracy a machine learning model can attain in predicting the lung condition of a IPF patient on their next 3 checkups? What method gives this accuracy?

Objectives

Short Term Objectives. Create 7 machine learning models (described in detail in Methodology) that predict the severity of a case of IPF with high accuracy, then combine their results using ensemble model methods.

Long Term Objectives. Create a graphical user interface for the program that medical professionals will be able to use to enter their base lung CT scan along with the FVC measurements to get an estimated severity along with a measure of confidence of the model. The model will have minimized the loss function to a global minimum.

Variables

Independent Variables. Since there are many ways to create such a model, there are any independent variables. The most important ones are the machine learning model used (the 7 methods used are listed in Methodology), how long the model trains (since the dataset is limited and there is fear of over-fitting), and the descent method (Adam or gradient descent) used by the model. These will affect the outcomes of the project.

Dependent Variable. The model's accuracy is the dependent variable. There are many measures of the model's accuracy, and one must be chosen to be used for all models to keep it consistent. The specific measure of accuracy will be chosen after analysing the data and the effects of the measures on the way the model trains.

Controlled Variables The only controlled variable is the data provided by Open Source Image Consortium containing CT scans and FVC measurements (*17*).

Methodology

Theoretical. Since the dataset contains images in the form of a base CT scan, the use of a convolutional neural network would be viable (*18*). Along with this, the use of certain protocols such as k-fold learning would streamline the training process (*19*). The use of ensemble learning could potentially increase the accuracy of the model and reduce the dangers of over-fitting and

under-fitting (*20*). The following are the methods that will be implemented into the ensemble: Linear Decay (*21*), Feature Engineering with a Linear Model (*22*), Extreme Gradient Boosting (*21*), Bayesian Learning (*23*), Auto-Encoder Training (*24*), as well as using Quantile Regression (*25*) with a Convolutional Neural Network (ResNet and EfficientNet) (*26*). These methods are all very different and will be a good test bed of models that will show their effectiveness. The ensemble (*27*) will combine all of these methods with correct weights in a neural network of its own.

Along with the creation of this ensemble model, it is possible to use machine learning to segment the lung CT scans into sections that would contain the information relevant to give an accurate prognosis. Although this is not directly related to creating an algorithm to determine the severity of the disease, this would be useful information for doctors to have when examining CT scans of their patients. Doctors would be able to determine the sections of the lungs responsible for a more severe case of IPF.

Implementation. In the short term, the data given in the kaggle competition (*17*) will be sorted and analysed to get an understanding of what measure of error would be best. The machine learning toolkit used relies heavily on python, and the models will be coded in python (*28*). Tensor Flow will be used for the vector and matrix manipulation, and for the creation of the models (*29*). These will be hosted locally or on a cloud computing server to reduce computation time. By creating the models individually, then calling all the methods in the ensemble method, the model will output one prediction.

Significance

Is is important that a simple, effective prognosis for IPF is found as it would reduce the work of doctors and help patients alike. Since IPF affects greater than the 0.005% of individuals that

was previously hypothesized (*4*), it is important to know more about this disease where more and more cases are being found.

Current prognosis methods for IPF use machine learning such as in (*11*), but the output of the model is a categorical value, rather than a numerical prediction for the forced vital capacity of the patient's lungs. Other methods also use data harder to obtain such as the measurement of radiological changes in (*10*). For this reason, a machine learning model of high accuracy based on one initial lung CT scan and FVC values on subsequent doctor visits would make the work of doctors much easier and would require scans which aren't as time consuming.

If successful, a model with great accuracy would make it such that patients with IPF would know the severity of their disease. This could reduce the stress and anxiety of patients.

Furthermore, a more accurate prediction could help doctors test potential cures of IPF. The model's predictions could act as a metric to measure if the cure truly works. If the cure changes the predictions of the neural network over a certain time period, then it is likely that the cure is effective.

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

18. L. Xu, J. S. Ren, C. Liu, J. Jia, *Advances in neural information processing systems* (2014), pp. 1790–1798.

19. J. Friedman, T. Hastie, R. Tibshirani, *The elements of statistical learning*, vol. 1 (Springer series in statistics New York, 2001).

20. T. G. Dietterich, *et al.*, *The handbook of brain theory and neural networks* **2**, 110 (2002).

21. T. Chen, C. Guestrin, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (Association for Computing Machinery, New York, NY, USA, 2016), p. 785–794. doi:10.1145/2939672.2939785.

22. A. Zheng, A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* (O'Reilly Media, Inc., 2018), first edn.

23. R. M. Neal, *Bayesian Learning for Neural Networks* (Springer-Verlag, Berlin, Heidelberg, 1996).

24. P. Baldi, *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11 (JMLR.org, 2011), p. 37–50.

25. Y. Jia, J.-H. Jeong, Deep learning for quantile regression: Deepquantreg (2020).

26. L. C. Jain, L. R. Medsker, *Recurrent Neural Networks: Design and Applications* (CRC Press, Inc., USA, 1999), first edn.

27. T. G. Dietterich, *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00 (Springer-Verlag, Berlin, Heidelberg, 2000), p. 1–15.

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

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