Prognosing Idiopathic Pulmonary Fibrosis with Machine Learning

Arnav Kumar

1. Introduction

Idiopathic Pulmonary Fibrosis (IPF) is a disease affecting the lung base which leads to lung function decline and has little to no therapies available other than lung transplant (Gross and Hunninghake, 2001; Mason et al., 1999). The disease affects more than 5 out of every 100,000 individuals (Coultas et al., 1994; Mason et al., 1999; Raghu et al., 2018). IPF is age-related, and has a median diagnosis age of 66 but there is no established cause (King Jr et al., 2011; Raghu et al., 2018). Patients of IPF experience a shortness of breath, and exhaustion after light excercise (Gross and Hunninghake, 2001). The outcome of Pulmonary Fibrosis can range from a healthy stability to a rapid health declination and eventually death (Robbie et al., 2017). Doctors are unable to easily diagnose disease severity as existing maethods are complicated, time consuming and are not standardized (Robbie et al., 2017).

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.

Machine learning is a good fit for the task at hand because of its impartiality, and its prior use for disease diagnosis and prognosis (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 measurements required are difficult to obtain or the disease severity is categorized rather than numerically predicted (Walsh et al., 2018).

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, thus predicting the rate at which the lung condition degrades. The main question of interest is which machine learning model produces the greatest accuracy in predicting the FVC of a IPF patient on their next 3 checkups, and is most suitable for use in the medical field.

2. Procedure

This study employs the use of many machine learning models, some of which are modified and influenced from the work of others (Consortium, 2020). 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). The data for this project is provided by the Open Source Imaging Consortium (Consortium, 2020). Exploratory data analysis was performed, and the data was preprocessed for use by the models. The data was split into two categories, the training data which all the models trained on, and testing data which the models had never seem before.

The **linear regression (LR)** method relies on the assumption that the FVC can be expressed as a linear combination of the input features. From every patient's first checkup, `weeks_passed` and `first_FVC` features were added to the data. The Scikit-learn package was then used to create a linear regressor which was then trained, and the model accuracy was measured.

Several **Dense Neural Networks (DNN)** were created, each with a different architecture. The data was first formatted in the same way used for linear regression. The models were then trained on the training data, the model with the most accurate predictions was chosen, and the model accuracy was calculated.

The base **Auto-encoder** utilized in this study was created by Welf Crozzo (Crozzo, 2020a). The encoder was used to stride over the data, adding 2000 extra features based on the patient's CT scan images. Using this new input data, a linear regression and many simple neural network models were created, which were then trained on the training data, and used to predict the FVC for the testing data. The best simple neural network was selected and the model accuracies were calculated for the linear regression and simple neural network models.

The **Bayesian Partial Pooling** method was modified from Carlos Souza (Souza, 2020). Features were removed from the data, and the data was reformatted into a matrix completion task. The partial pooling bayesian hierarchical model was created and trained. The model was used to predict the FVC for the testing data, and the model accuracy was calculated.

The **Linear Decay** method used here is a modification of Welf Crozzo's work (Crozzo, 2020b). The data was first formatted and some patients with poor CT scans were removed, then a linear decay model was created with a convolutional neural network to predict the unknown model coefficients. The convolutional neural network was trained and the model coefficients were found, and the model accuracy was then calculated.

The accuracy of a model is measured using its Laplace Log Likelihood (LLL). The model's FVC prediction, the true FVC, and the model's confidence are required to calculate the LLL. A LLL closer to 0 represents a model which is more accurate, but the score 0 itself is unattainable for all practical purposes, instead, an impressive score would be around -6.5. The worst score a model should get is -8.023, and any model with a LLL lower than -8.023 is useless (Rao, 2020).

3. Results

Figure 1 displays the model performance of the models analysed using the model's Laplace Log Likelihood. Out of the testing data (which the model has never seen during training), 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 metric values for the private and public testing data due to GPU time limits. Both models take over 4 hours to run, and hence could not be evaluated.

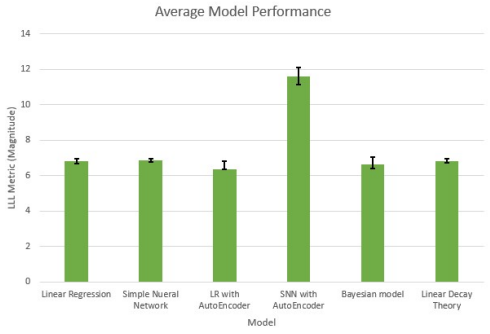
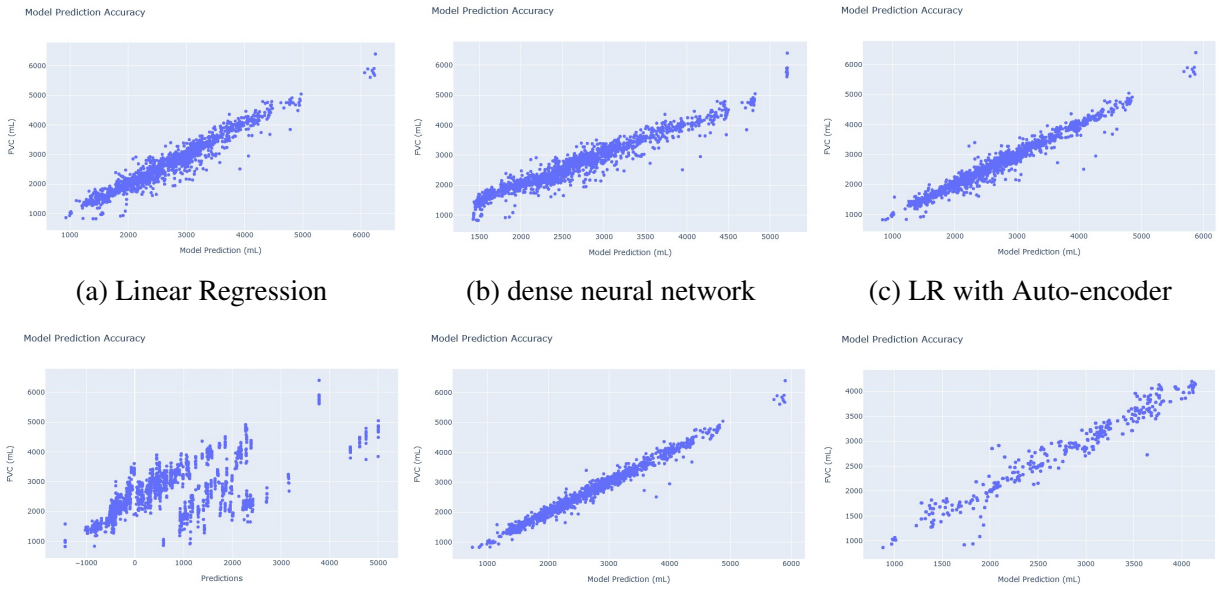


Figure 1: Comparison of Average Model LLL

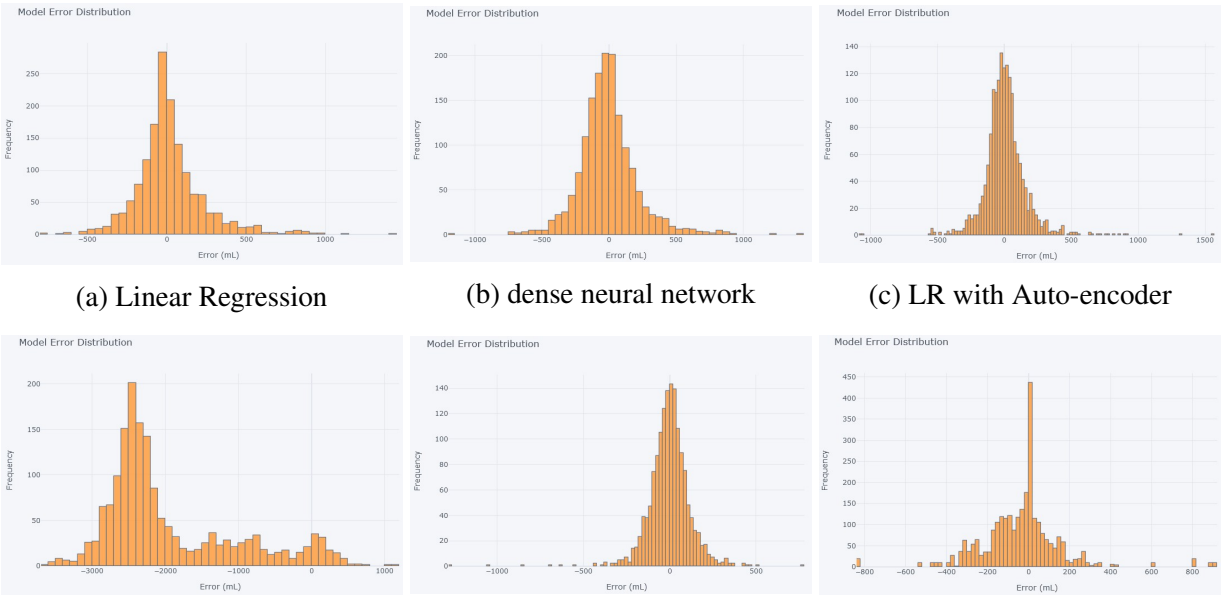
Figure 2 shows the accuracy of the predictions of several models. The true patient FVC is graphed against the model prediction, so a scatterplot closer to the line $y = x$ means the model is more accurate. In addition, Figure 3 is a histogram of the errors of the models and we desire an error which has low spread, is unimodal, and is centered at 0.



(a) Linear Regression (b) dense neural network (c) LR with Auto-encoder

(d) DNN with Auto-encoder (e) Bayesian Partial Pooling (f) Linear Decay

Figure 2: Plots of True FVC vs Model Prediction



(a) Linear Regression (b) dense neural network (c) LR with Auto-encoder

(d) DNN with Auto-encoder (e) Bayesian Partial Pooling (f) Linear Decay

Figure 3: Model Error Distributions

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 as in Figure 1.

Overall, models which use neural networks performed poorly. Introducing extra layers and a lot or more tunable variables, decreases the chance that the model will find the optimal combinations of weights. Instead, the model will likely end up with a suboptimal set of weights and biases, and will have reached a local minimum rather than a global minimum.

We see this with the Dense Neural Network with Auto-encoder features which has a LLL score worse than -8.023. Its poor performance can be attributed to the high number of input features of the model, and hence the number of tunable weights and biases.

Another interesting find is that the Bayesian Partial Pooling method seemed to overfit the training data as its performance on the training data was much better than the testing data.

Furthermore, Figure 3f seems to suggest that the Linear Decay Method performed poorly, but it's irregular error distribution is actually due to the fewer patients the Linear Decay trained on. The model's accuracy on the testing data showed no compromise.

Overall, use of the Linear Decay Theory model is advised for its accuracy, consistency, and useful confidence values. For the field of medicine, having a method which is well understood is preferred, and statistical methods such as the Linear Decay Theory model are guaranteed to always perform as expected.

The use of the Linear Decay Theory Model would not only eliminate human bias in the prognosis and make it easier for medical professionals, but it would give patients enough time to come to terms with their disease.

Additionally, the lesson of avoiding overly complex models can be applied to other projects and has been described by Occam's razor, which states that when there are multiple competing hypothesis (the multiple models being compared), the hypothesis with the simplest assumption (the assumption that FVC is a linear function of features) is the best hypothesis.

5. Acknowledgements

Thank you to Dr. Christian Jacob for supporting and guiding me through the project and to my teachers, Dr. Beatriz Garcia-Diaz and Ms. Bogusia Gierus for their continued support. Finally, thank you to Mr. Chuck Buckley for providing invaluable feedback.

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., ... 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

du Bois, R. M., Albera, C., ... 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

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

Maldonado, F., Moua, T., ... 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, ... 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., ... 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.

Pedregosa, F., Varoquaux, G., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Raghu, G., Remy-Jardin, M., … 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

Rao, R. (2020). Osic: Understanding laplace log likelihood. https://www.kaggle.com/rohanrao/osic-understanding-laplace-log-likelihood

Robbie, H., Daccord, C., … 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.

Walsh, S. L., Calandriello, L., … 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

Wang, Y., Fan, Y., … 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