The 12th IEEE International Conference on Dependable Systems, Services and Technologies, DESSERT'2022
9-11 December, 2022, Athens, Greece

A regression-based machine learning approach for the prediction of lung function decline

Angeliki Poulou
Laboratory of Information Technologies
Faculty of Information Science and Informatics
Ionian University
angeliki.p300@gmail.com

Marios Poulos
Laboratory of Information Technologies
Faculty of Information Science and Informatics
Ionian University
mpoulos@ionio.gr

Maximilianos Panas
Laboratory of Information Technologies
Faculty of Information Science and Informatics
Ionian University
m@maxpanas.com

2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT) | 979-8-3503-3304-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/DESSERT58054.2022.10018624

***Abstract*—Pulmonary fibrosis is a progressive disease of the lungs which usually gets worse over time. Once this disease damages the lungs, it cannot be cured totally, but early detection and proper diagnosis can help to keep the disease in control. The Kaggle competition entitled "OSIC Pulmonary Fibrosis Progression Predict lung function decline" ran from July to September 2020 with the goal of early detection of the disease. Our approach achieved a Laplace Log Likelihood score of −6.8590 which was within the bronze medal band. The Kaggle dataset contained CT scans and anonymized demographic and clinical data from multiple patient visits, such as spirometry forced vital capacity (FVC), for 176 unique patients. In our method we predict FVC and a confidence measure using a sigmoid equation. This equation is extracted via a novel transformation using only three of the given parameters. In this way we created a simple but accurate model for the prediction of lung function decline.**

Keywords—Machine Learning; Regression algorithms; Prediction error; Pulmonary Fibrosis; Prognostic tool.

I. INTRODUCTION

The natural history of fibrotic interstitial lung disease is variable. The worst prognosis is when idiopathic pulmonary fibrosis progresses toward respiratory failure, occurring on average about 4 years after initial diagnosis. Median survival for progression of this disease varies between 2 and 5 years [1]. There is great variability in the disease's course for each patient, which depends on factors such as the occurrence of acute exacerbations during its course and the presence of other illnesses. Results can range from long-term stability to rapid deterioration, but doctors can't easily tell where a person falls on this spectrum. Computational analysis of patient data can contribute to this prediction, which can greatly help both patients and physicians. With current methods, fibrotic lung disease is difficult to treat, even with access to chest CT scans [2,3]. In addition, the widely varying prognosis poses problems in organizing clinical trials. Finally, in addition to fibrosis-related symptoms, patients are known to suffer from extreme anxiety due to the unclear course of the disease.

The Open-Source Imaging Consortium (OSIC) [4] is a non-profit organization spanning academia, industry, and philanthropy. Its mission is to combat respiratory diseases, including idiopathic pulmonary fibrosis, interstitial lung disease fibrosis, and emphysematous disease, by bringing together radiologists, clinicians and computer scientists to improve diagnosis through image processing and data analysis of sourced patient scans and clinical data.

The competition ran by OSIC on Kaggle, provided CT scans of patient lungs along with additional anonymized patient data and had a goal to predict the extent of deterioration of lung function at future time-points from the patient's initial consultation using machine learning techniques [5]. Forced vital capacity (FVC) based on spirometer readings, which measures how much air is inhaled and exhaled, was used to determine lung function. With performant and acceptably accurate predictions, patients and their families can have a better understanding of prognosis when first diagnosed with this incurable lung disease. Improved severity detection also has the potential for a positive impact on therapeutic study design, possibly accelerating clinical development of new therapies.

Our application scenario is to leverage patient spirometry measurements to predict FVC at a desired time-point in the future to aid clinicians with treatment and care planning and is complementary to continuous monitoring through spirometry tests. In our approach, we predict future FVC and a confidence measure for a patient using a sigmoid equation. This equation is extracted via a novel transformation using only three of the available patient parameters. The novelty of our model lies in requiring only FVC measurements and avoiding the need to interpret the baseline CT scan, thus dramatically reducing the computational complexity of the prediction without sacrificing significant accuracy over the given test set.

II. MATERIALS & METHODS

Materials: To construct the proposed fibrosis prognostic model, we used the patient cohort from the OSIC Pulmonary Fibrosis Progression Challenge (OSIC, 2020) [5]. The data for this patient cohort consists of anonymized chest CT scans, FVC measurements from frequent visits over the course of 1–2 years, and associated clinical metadata (i.e., age, sex, smoking status, and patient's relative FVC measurement compared to the typical FVC measurement of a patient with similar characteristics). The CT scan for each patient was acquired at Weeks = 0 in the dataset and was accompanied by about 9 spirometry measurements taken a given number of weeks pre/post this baseline CT scan. For the competition, the

dataset of scans and clinical data was split 15% and 85% between public and private datasets respectively, with the public set containing scans and clinical data for 176 patients. Whereas the anonymized training data contained the baseline CT scan and the entire history of FVC measurements, the test set contained only the CT scan and an initial FVC measurement, based on which the prediction should be made.

In our model, after experimentation, three clinical parameters are used per patient: Weeks (w_i), FVC (f_i) and Percent (p_i), where:

- w_i is the relative number of weeks pre/post the baseline CT scan (may be negative),
- f_i is the recorded lung capacity in ml (FVC) on w_i and,
- p_i is a computed field which approximates the patient's FVC on w_i as a percent of the typical FVC for a person of similar characteristics (sex, smoking status, etc.).

In the training dataset, we observed that the ratio between f_i and p_i values is constant and unique per patient:

$$\frac{f_i}{p_i} = C \tag{1}$$

Methods: From the three aforementioned parameters, the following variables are used for our training and analysis:

- w_1 is the Weeks since the baseline CT scan, at which point the known spirometry measurement was collected
- w_x is the Weeks since the baseline CT scan, for which point we want to predict the FVC
- f_1 is the known FVC at w_1
- f_x is the resulting predicted FVC at w_x
- p_1 is the known Percent value at w_1

TABLE I. SAMPLE PATIENT CLINICAL DATA

i	*Weeks (w_i)*	*FVC (f_i) in ml*	*Percent (p_i)*	*FVC / Percent (C)*
1	17	3294	79.2589	41.5600
2	18	2777	66.8191	41.5600
3	19	2700	64.9663	41.5600
4	21	3014	72.5217	41.5600
5	23	2661	64.0279	41.5600
6	30	2778	66.8431	41.5600
7	42	2516	60.5390	41.5600
8	53	2432	58.5178	41.5600
9	70	2578	62.0308	41.5600

Clinical records for an example patient, ID00423637202312137826377, from the OSIC Kaggle training dataset.

As part of our analysis, we applied a data transformation to calculate the difference (d) of the angles determined by the vectors (f, w) and (f, p), making use of the linear relationship between f_i and p_i for each patient, Eq. (1), and expressing $p_x = f_x/C$ [6]:

$$d_x = \arctan\left(\frac{f_x - f_1}{\frac{f_x}{C} - p_1}\right) - \arctan\left(\frac{f_x - f_1}{w_x - w_1}\right) \tag{2}$$

and calculated the slope (s) between f and w:

$$s_x = \frac{f_x - f_1}{w_x - w_1} \tag{3}$$

Equations (2) and (3) were then applied to 172 patients from the training data that each have around 9 FVC readings (see Table 1) collected over a period of approximately 1–2 years. Specifically, for each patient, w_1 was set to the first given week of the training data and w_x was varied for all their available follow-up records after w_1. This resulted in about 1376 s and d pairs. By plotting s with respect to d using these 1376 pairs, Fig. (1), we notice a sigmoid curve.

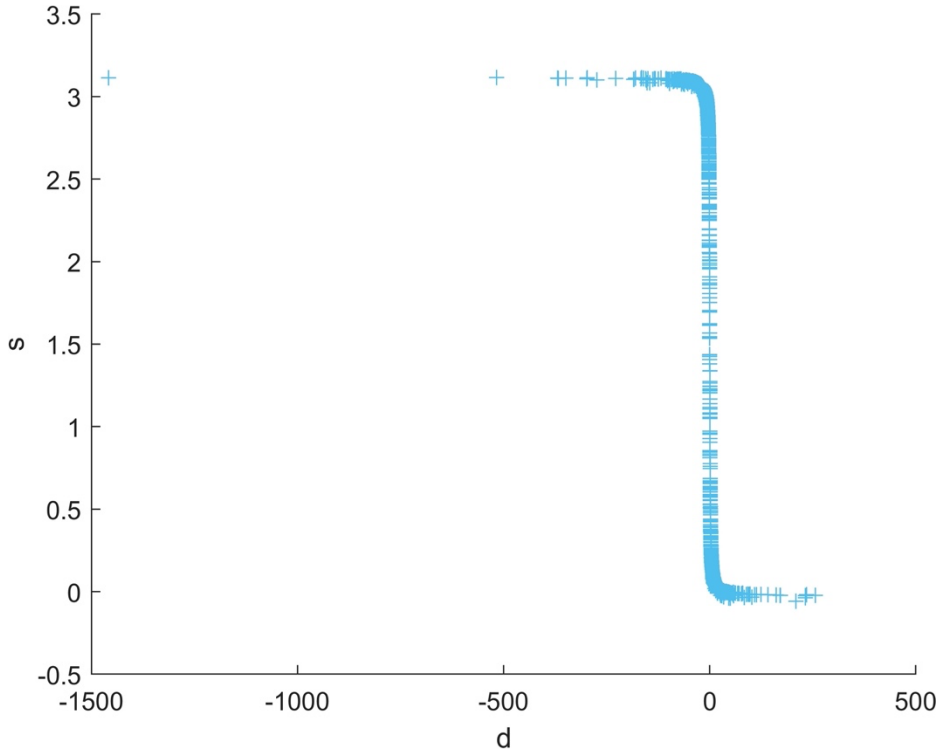

Fig. 1. Plot of s with respect to d.

III. TRAINING RESULTS

In the training procedure of our model the 1376 sets of given parameters (f_1, p_1, w_1, w_x) were used to fit a curve using an optimization of the parameters of the following sigmoid function [7, 8]:

$$d(s) = k + \frac{l - k}{10^{(m-s)}n + 1} \qquad (5)$$

The extracted parameters of this fitting procedure are:

$$\begin{aligned} k &= 3.0483 \\ l &= 0.0357 \\ m &= -0.0035 \\ n &= 0.4015 \end{aligned}$$

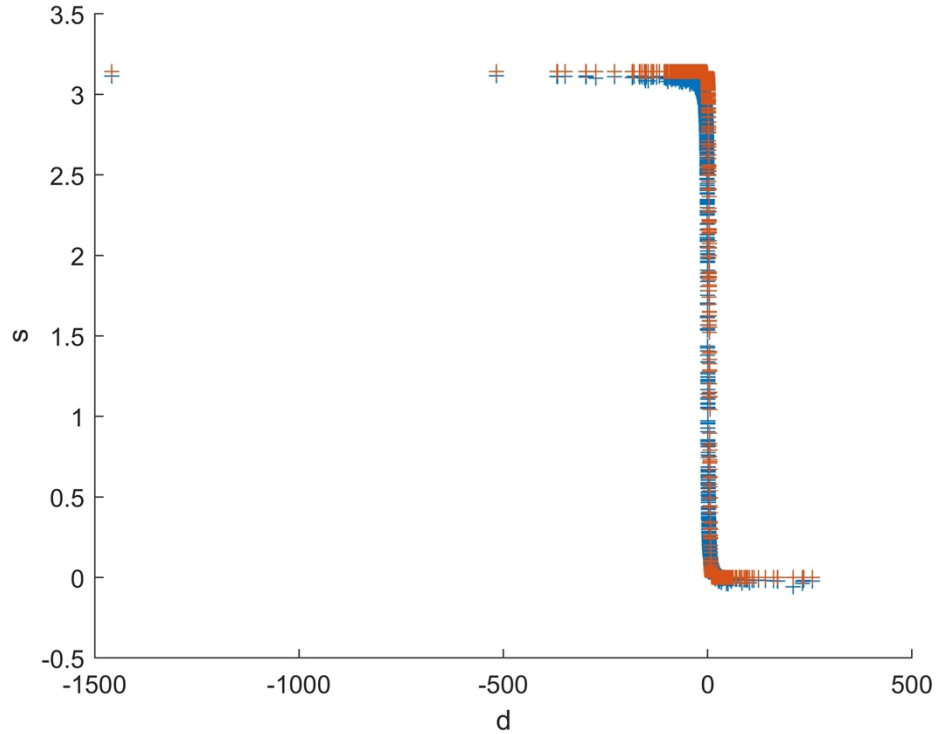

Fig.2. The fitted results (in red) plotted together with the experimental d and s data points (in blue).

The above sigmoid function describes the average relationships slightly better than the linear function [9, 10], with a R-squared measure of goodness of fit of >0.05 considered suitable for the test. In Fig. (2) the resulting s and d points of this fitting procedure (in red) are plotted together with the experimental data points (in blue).

This sigmoid relationship between d and s was found, via experimentation, to also be described by the following expression:

$$d(s) = \frac{f_1}{p_1 + e^{-s}} \tag{6}$$

with the solution to its second derivative with respect to s being $s = -\log(p_1)$. By combining this root with Eq. (3), the following expression for f_x with respect to w_1, w_x and p_1 can be written:

$$f_x = f_1 - \log(p_1) \cdot (w_x - w_1) \tag{7}$$

This expression can thus also be used to predict the FVC value at a future week w_x for any initial values p_1 and f_1 given for a patient on w_1.

IV. Evaluation & Discussion

The OSIC Kaggle competition was evaluated on a modified version of the Laplace Log Likelihood [11] method. In medical applications, it is useful to evaluate a model's confidence in its decisions. Accordingly, the error metric used is designed to reflect both the accuracy and certainty of each prediction.

$$S_{\text{clipped}} = \max(S, 70) \tag{8}$$

$$D = \min(|f_{\text{true}} - f_x|, 1000) \tag{9}$$

$$\text{metric} = -\frac{\sqrt{2}D}{S_{\text{clipped}}} - \ln\left(\sqrt{2}S_{\text{clipped}}\right) \tag{10}$$

The error has a threshold prediction difference (D) at 1000 ml to avoid large errors adversely penalising results, while the confidence values (S) are clipped at 70 ml to reflect the approximate measurement uncertainty in FVC [12]. The final error score is calculated by averaging the metric across predictions for the final three (hidden) FVC measurements for all patients in the test set. In the testing procedure, we investigated 1376 cases in the metric calculation via equations 8–10.

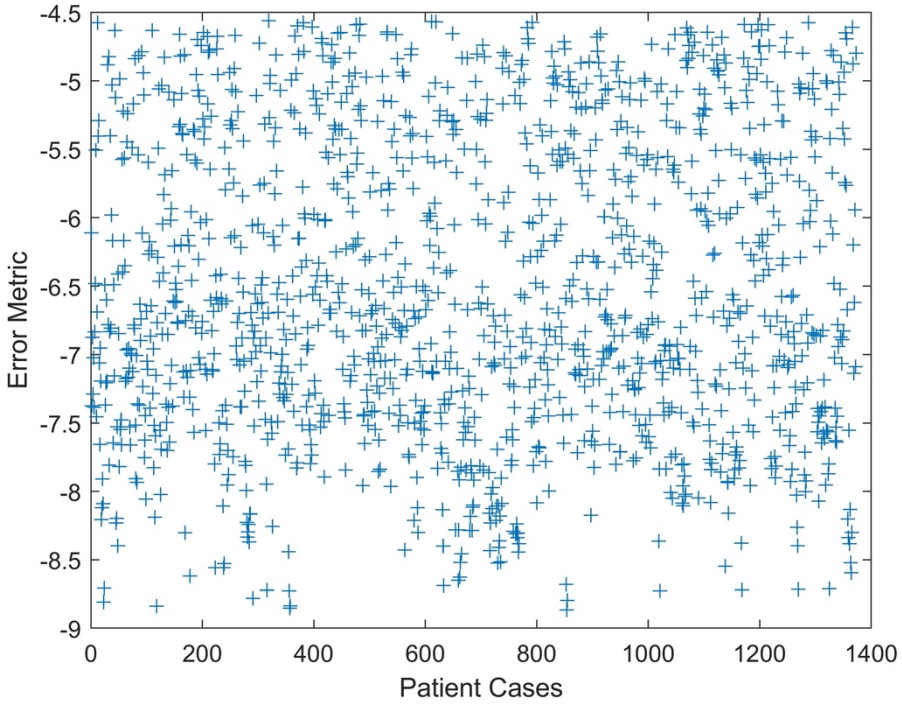

Fig.3. Error score for each of the patient records when applying Eq. (10).

We calculated the mean error score on the public dataset to be −6.5587 and −6.5187 for the fitted sigmoid equation, Eq. (5), and Eq. (6) respectively, indicating that the latter has better predictive behavior. The error values for the model based on Eq. (6) are presented in Fig. (3). Furthermore, this best performing model was further evaluated during the final stage of the OSIC competition using the private dataset consisting of 7797 patient records, 85% of the total dataset, yielding an error score of −6.8590.

In Table (2), a comparison of our model with other state-of-art models published in the literature as well as other models publicly available from the OSIC Kaggle competition are shown, ordered based on their Laplace Log Likelihood scores. Based on this metric, the predictive performance of our model is higher compared to other methods that also avoid CT scan image processing but use more clinical features [13]. Methods that include CT scan data in their model, denoted in the table with a *, on the other hand do generally outperform our approach with error scores between −6.641 and −6.8305 [2,14,15,16]. It is noted however that the Multiple Quantile Regression method [2], has lower performance, with an error

score of −6.92, than our method despite incorporating CT scan image data in its predictions.

TABLE II. METHOD COMPARISON

Comparison with different methods	***Laplace Log Likelihood***
FVC-Net* [15]	−6.641
Elastic Net Regression* [2]	−6.73
Ridge Regression* [2]	−6.81
Fibrosis Net* [16]	−6.8188
Kaggle 1st place* [14]	−6.8305
Our Solution	−6.859
Simple LogReg [13]	−6.8648
Multiple Quantile Regression* [2]	−6.92

* denotes studies that have used CT scan image processing for their prediction.

V. CONCLUSIONS

In summary, we developed a regression-based model via parameter optimization on a sigmoid function to predict future FVC based on historical patient data that yielded good predictive accuracy on the OSIC Kaggle competition on Pulmonary Fibrosis and predicting lung function decline. This model, trained on data from 172 patients, achieves a Laplace Log Likelihood score of −6.8590 on the private dataset consisting of over 1000 patient records used in the competition. A key importance of this model is that it produces accurate predictions of future FVC while avoiding the computationally complex evaluation of CT scan images of patient lungs. Good and timely predictive ability for lung function decline is crucially important for early identification and treatment of patients with pulmonary fibrosis. In the future this methodology can serve as a basis for further improvement in prediction of lung function decline based on historical patient data.

REFERENCES

[1] Spagnolo, P., Ryerson, C. J., Putman, R., Oldham, J., Salisbury, M., Sverzellati, N., ... & Cottin, V. (2021). Early diagnosis of fibrotic interstitial lung disease: challenges and opportunities. *The Lancet Respiratory Medicine*, *9*(9), 1065-1076.

[2] Mandal, S., Balas, V. E., Shaw, R. N., & Ghosh, A. (2020, October). Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In *2020 IEEE International conference on computing, power and communication technologies (GUCON)* (pp. 861-865). IEEE.

[3] Morozov, S. P., Chernina, V. Y., Blokhin, A. I., & Gombolevskiy, V. A. (2020). Chest computed tomography for outcome prediction in laboratory-confirmed COVID-19: A retrospective analysis of 38,051 cases. *Digital Diagnostics*, *1*(1), 27-36.

[4] https://www.osicild.org, Retrieved December 4, 2022

[5] https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression, Retrieved December 4, 2022

[6] Schatte, P. (1999). Computing the Angle between Vectors. *Computing*, *63*(1).

[7] Pavão, R., Savietto, J. P., Sato, J. R., Xavier, G. F., & Helene, A. F. (2016). On sequence learning models: Open-loop control not strictly guided by Hick's law. Scientific reports, 6(1), 1-9.

[8] Castro-Neto, M., Jeong, Y., Jeong, M. K., & Han, L. D. (2009). AADT prediction using support vector regression with data-dependent parameters. *Expert Systems with Applications*, *36*(2), 2979-2986.

[9] Meddings, J. B., Scott, R. B., & Fick, G. H. (1989). Analysis and comparison of sigmoidal curves: application to dose-response data. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, *257*(6), G982-G989.

[10] R P (2022). sigm_fit (https://www.mathworks.com /matlabcentral/ fileexchange/42641-sigm_fit), MATLAB Central File Exchange. Retrieved December 4, 2022.

[11] Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, *4*(1), 12-35.

[12] Kittler, J., & Illingworth, J. (1986). Minimum error thresholding. *Pattern recognition*, *19*(1), 41-47.

[13] https://www.kaggle.com/code/artkulak/simple-logreg/notebook?scriptVersionId=44081090, Retrieved December 4, 2022.

[14] https://www.kaggle.com/competitions/osic-pulmonary-fibrosis-progression/discussion/189346, Retrieved December 4, 2022.

[15] Yadav, A., Saxena, R., Kumar, A., Walia, T. S., Zaguia, A., & Kamal, S. M. M. (2022). FVC-NET: An Automated Diagnosis of Pulmonary Fibrosis Progression Prediction Using Honeycombing and Deep Learning. *Computational Intelligence and Neuroscience, 2022*, 2832400.

[16] Wong, A., Lu, J., Dorfman, A., McInnis, P., Famouri, M., Manary, D., Lee, J. R. H., Lynch, M. (2021). Fibrosis-Net: A Tailored Deep Convolutional Neural Network Design for Prediction of Pulmonary Fibrosis Progression From Chest CT Images. *Frontiers in Artificial Intelligence, 4*, 161.