Pulmonary Fibrosis Progression Prognosis Using Machine Learning

Arseny Glotov
Department of Mathematical Modeling
Institute of Mathematics and Information Technologies (named after Prof. Nikolay Chervyakov)
North Caucasus Federal University
Stavropol, Russia
senya.ds.researcher@gmail.com

Pavel Lyakhov
Department of Mathematical Modeling
Institute of Mathematics and Information Technologies (named after Prof. Nikolay Chervyakov)
North Caucasus Federal University
Stavropol, Russia
ljahov@mail.ru

2021 IEEE Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT) | 978-1-7281-7691-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/USBEREIT51232.2021.9455070

***Abstract*—Lung fibrosis means scarring of tissue in a patient's lungs and is a common condition that can complicate the course of COVID-19 disease. Pulmonary fibrosis destroys the patient's lungs, preventing oxygenation of the blood. Modern methods of treatment are not highly effective even with access to a patient's CT scan. The problem of predicting the prognosis of pulmonary fibrosis is extremely important, since its solution will make it possible to organize clinical trials to study methods of treating patients with fibrosis more effectively. This article proposes a method for predicting the prognosis of pulmonary fibrosis progression as the volume of inhaled and exhaled air to the individual patient based on tabular patient data using an ensemble of four machine learning algorithms. This solution also provides a forecast accuracy because it is useful in medical applications to assess the "confidence" of the model in its predictions. Modeling the proposed method shows a better result than other forecasting methods that are compared in the article.**

Keywords—Pulmonary fibrosis progression prognosis, Machine learning, Computer-aided diagnostics.

I. Introduction

Pulmonary fibrosis (PF) is a state in which lung tissue gets scarred. This scar tissue then interferes with breathing, and is the reason of reducing tissue integrity which in turn complicates the passage of oxygen through the alveoli (bubbles in which air communicates with blood). In this case, normal lung tissue is replaced by connective tissue. The process of tissue regeneration involving return to the original state of the lung is not possible, so the patient may not fully recover, but his condition can still be managed [1].

Pulmonary fibrosis, however, does not have to be caused by a certain detectable pathogen or is defined by a diagnosed initial acute inflammatory phase. As a rule, it can be connected to a severe lung injury. The condition is typically caused or accompanied by respiratory infections, chronic granulomatous disease, medication, and connective tissue disorders. According to current clinical, radiographic, and autopsy data, pulmonary fibrosis is the main reason for severe acute respiratory distress syndrome (SARS) and MERS pathology, and up-to-date research allows to conclude that PF can also enhance SARS-CoV-2 infection [2].

Modern methods do not allow full treatment of pulmonary fibrotic diseases, even with access to a computed tomography scan of the chest. In addition, the wide range of different prognoses poses challenges when organizing clinical trials. Finally, patients suffer from severe anxiety - in addition to symptoms associated with fibrosis - due to the opaque pathway of disease progression [3].

In this article, we present a solution to the OSIC Pulmonary Fibrosis Progression problem from the Open Source Imaging Consortium on the Kaggle platform [3]. Our approach uses a combined forecast of four algorithms: DNN (Deep Neural Network), GBDT (Gradient Boosting Decision Tree) [4], NGBoost (Natural Gradient Boosting) [5] and ElasticNet [6]. This solution provides a prediction of lung function as the volume of inhaled and exhaled air to the individual patient, as well as the accuracy of the forecast.

II. Formulation of the Problem

A. Dataset

The dataset provides basic computed tomography of the chest and related clinical information for a group of patients in tabular form (Table 1). There are 176 cases in the train set and around 170 cases in the test set (test set is hidden, and it was used to check the final accuracy of the algorithm on the Kaggle server).

TABLE I. Structure of the Tabular Part of the Dataset

Field name	Description
Patient	Patient's unique identifier
Weeks	Relative number of weeks before / after baseline CT (may be negative)
FVC	Recorded lung capacity in ml
Percent	A calculated value that roughly corresponds to the percentage of the typical FVC for a healthy person with similar parameters
Age	Patient's age
Sex	Patient's gender
SmokingStatus	Patient's smoking status: currently smokes, ex-smoker, never smoked

Baseline CT of a patient was performed at time *Weeks = 0* and then the patient's forced vital capacity (FVC) was measured over a period of approximately 1-2 years.

Since this is real medical data, the relative time to measure the forced vital capacity of the lungs is very different. Initial measurement time relative to CT and duration to predicted points may be different for each patient [7].

B. Evaluation

Solutions to this problem were evaluated using a modified version of Laplace Log Likelihood [8]. It is useful to assess

This work was supported by the Russian Foundation for Basic Research (project no. 19-07-00130 A) and also by the Presidential Council for grants (project no. MK-3918.2021.1.6).

the “confidence” of the model in its predictions when solving applied medical problems. Consequently, this metric allows us to get an estimate of both the forecast accuracy and the level of confidence of the algorithm in it.

For each true FVC measurement, both the FVC and the confidence score (standard deviation σ) had to be predicted. The metric is calculated as:

$$\sigma_{\text{clipped}} = \max(\sigma, 70), \tag{1}$$

$$\Delta = \min\left(\left|\text{FVC}_{\text{true}} - \text{FVC}_{\text{predicted}}\right|, 1000\right), \tag{2}$$

$$\text{metric} = -\frac{\sqrt{2}\Delta}{\sigma_{\text{clipped}}} - \ln\left(\sqrt{2}\sigma_{\text{clipped}}\right). \tag{3}$$

Error threshold is 1000 ml (2) to avoid large errors that adversely affect results, while confidence values are limited to 70 ml (1) to reflect the approximate uncertainty of FVC measurement. The final metric value is calculated by averaging it over the entire set of test data.

III. Proposed Method of Prognosis

Often solutions to such machine learning problems are based on large and varied ensembles, which is not always possible and feasible in real applications. During testing, we often want to minimize the amount of memory consumed and the time of the prediction. In this article, we propose a solution consisting of four lightweight models trained exclusively on tabular patient information.

1) Deep Neural Network: the network architecture is shown in Fig. 1. The input is preprocessed tabular information about the patient, and the result of the neural network is a vector consisting of three elements $[x_1, x_2, x_3]$. The final network forecast is formed as follows:

$$\text{FVC}_{\text{predicted}} = x_2, \tag{4}$$

$$\sigma = x_3 - x_1. \tag{5}$$

The neural network training process is carried out through the optimization of the loss function that consists of two parts: the initial metric (3) and Pinball loss [9], which is defined as:

$$L_\tau(y, z) = \begin{cases} (y - z)\tau, & y \geq z \\ (z - y)(1 - \tau), & z > y' \end{cases} \tag{6}$$

where τ is the target quantile, y and z are the real value and the quantile forecast respectively. When solving the problem, we suggest to calculate quantiles of 0.2, 0.5, and 0.8. The final view of the loss function:

$$\text{Loss} = 0.4 \cdot L_\tau + 0.6 \cdot \text{metric}. \tag{7}$$

2) GBDT: we propose to use LightGBM [10] implementation of this algorithm. For training the model, a modified Gaussian Negative Log Likelihood Loss function is proposed:

$$\text{Loss}(t, \mu, \sigma) = \frac{(t-\mu)^2}{2\tilde{\sigma}^2} + \ln\left(\sqrt{2\pi}\tilde{\sigma}\right), \tag{8}$$

$$\tilde{\sigma} = \ln(1 + e^{\sigma}), \tag{9}$$

where t and μ are FVC_{true} and $\text{FVC}_{\text{predicted}}$ respectively.

3) NGBoost: probabilistic prediction (or probabilistic forecasting), defined as an approach where the model outputs a full probability distribution over the entire outcome space, is a natural way to quantify uncertainties of given predictions [4]. We propose to use the difference between the upper and lower bounds of the 10% confidence interval of the predicted FVC value as the predicted σ.

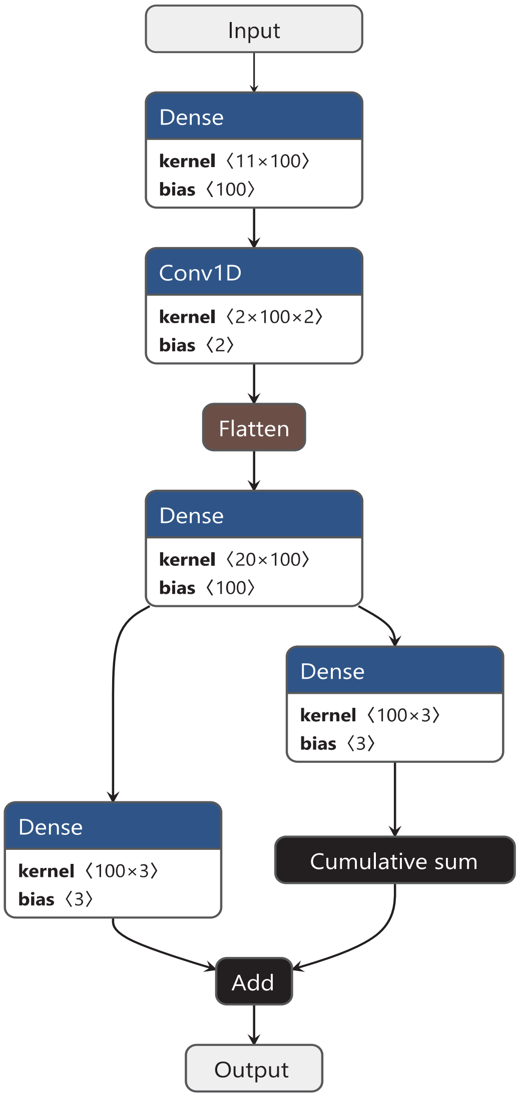

Fig. 1. Proposed deep neural network architecture for predicting pulmonary fibrosis progression.

4) ElasticNet: aims to combine the best of ridge regression and lasso regression by combining L_1 and L_2 regularization [6]. This model is proposed solely to reduce variance in the final prediction of FVC, therefore σ was not predicted.

IV. Simulation

For training and obtaining the final forecast, it is proposed to use the cross-validation algorithm on 5 folds with non-overlapping sets of patients [11]. On each fold, predictions were obtained for the test set as follows (Fig. 2):

1) By solving a system of linear algebraic equations using ordinary least squares method (OLS), the weights w_0, w_1, w_2, w_3 for the FVC predictions of each of the four algorithms were found [12]. The fold prediction is obtained by summing the FVC predictions of the four algorithms multiplied by their weights.

2) The values of σ were averaged.

We propose to average the FVC and σ calculated by the above prediction method over all folds to obtain the final prediction on a test dataset.

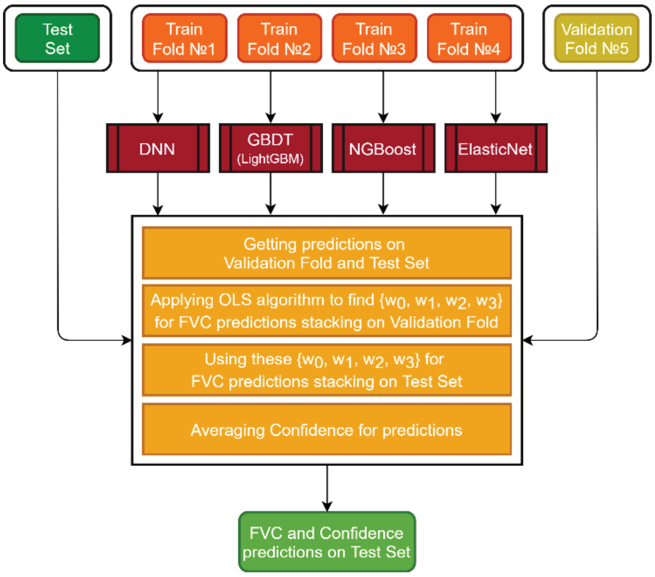
Test Set
Train Fold №1
Train Fold №2
Train Fold №3
Train Fold №4
Validation Fold №5
DNN
GBDT (LightGBM)
NGBoost
ElasticNet
Getting predictions on Validation Fold and Test Set
Applying OLS algorithm to find {w0, w1, w2, w3} for FVC predictions stacking on Validation Fold
Using these {w0, w1, w2, w3} for FVC predictions stacking on Test Set
Averaging Confidence for predictions
FVC and Confidence predictions on Test Set

Fig. 2. The scheme of training our presented method and getting predictions for each fold.

Table 2 shows the results of checking the accuracy of the algorithm by metric (3) for the test part of the data. The method proposed in this article for predicting the prognosis of pulmonary fibrosis shows the highest value of forecast accuracy in comparison with other considered algorithms.

TABLE II. RESULTS OF MODELING PREDICTIONS ON A TEST DATASET

Model		Metric
GBDT *(LightGBM)* [10]	**Known methods**	-6.8826
NGBoost [4]		-6.9362
ElasticNet *(σ = 250)* [6]		-7.9096
DNN (Fig. 1)		-6.8842
Simple averaging of predictions from DNN, GBDT, NGBoost and ElasticNet algorithms		-6.9607
Our proposed method		**-6.8507**

V. CONCLUSIONS

This article proposes an effective algorithm for predicting the pulmonary fibrosis progression. The algorithm is based on an ensemble of DNN, GBDT, NGBoost and ElasticNet. This approach makes it possible to use exclusively tabular patient data for forecasting without the need to access the results of his chest computed tomography. During the simulation, the proposed algorithm showed the best performance indicators in comparison with other considered methods. The code of the proposed method is available on Kaggle platform [13].

ACKNOWLEDGMENTS

The authors would like to thank the North Caucasus Federal University for supporting in the contest of projects competition of scientific groups and individual scientists of the North Caucasus Federal University.

REFERENCES

[1] "Pulmonary Fibrosis Overview," *Pulmonary Fibrosis Foundation*. [Online]. Available: https://www.pulmonaryfibrosis.org/life-with-pf/about-pf. [Accessed: 12-Jul-2020].

[2] A. S. Ojo, S. A. Balogun, O. T. Williams, and O. S. Ojo, "Pulmonary fibrosis in covid-19 survivors: Predictive factors and risk reduction strategies," *Pulmonary Medicine*, vol. 2020, 2020.

[3] "OSIC Pulmonary Fibrosis Progression," *Kaggle*. [Online]. Available: https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/overview. [Accessed: 07-Jul-2020].

[4] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[5] T. Duan, A. Anand, D. Y. Ding, K. K. Thai, S. Basu et al., "Ngboost: Natural gradient boosting for probabilistic prediction," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2690–2700.

[6] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[7] "OSIC Pulmonary Fibrosis Progression Data," *Kaggle*. [Online]. Available: https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/data. [Accessed: 07-Jul-2020].

[8] "OSIC Pulmonary Fibrosis Progression Evaluation," *Kaggle*. [Online]. Available: https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression/evaluation. [Accessed: 07-Jul-2020].

[9] G. Biau and B. Patra, "Sequential quantile prediction of time series," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1664–1674, 2011.

[10] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen et al., "LightGBM: A highly efficient gradient boosting decision tree", *Proc. NIPS*, pp. 1-9, 2017..

[11] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[12] B. Pavlyshenko, "Using stacking approaches for machine learning models," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 255–258.

[13] A. Glotov, "DNN + LGBM + NGBoost + ElasticNet," *Kaggle*, 25-Dec-2020. [Online]. Available: https://www.kaggle.com/benefactor/dnn-lgbm-ngboost-elasticnet. [Accessed: 25-Dec-2020].