

# Prediction Analysis of Idiopathic Pulmonary Fibrosis Progression from OSIC Dataset

Sampurna Mandal, Valentina E. Balas, Rabindra Nath Shaw and Ankush Ghosh

**Abstract**— Pulmonary fibrosis is a progressive lungs disease 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 this disease in control. It causes scarring in the lungs over time. As an effect, people face breathing difficulty. It can cause shortness of breath, even at rest. The general causes of pulmonary fibrosis can be exposure to toxic element like coal dust, asbestos fibres, silica dust, hard metal dusts etc. But in majority of the cases, the doctor cannot figure out the exact cause of this disease. That's why this disease is termed as Idiopathic Pulmonary Fibrosis. The objective of this paper is to analyse and compare the performance of various machine learning models by predicting the final forced volume capacity measurements for each patient and a confidence value. It can be deployed on any computer to predict a patient's severe condition regarding lungs function which is based on a CT scan of the lungs of the patients. Lung function is checked out based on a spirometer output that measures the forced vital capacity (FVC) of the lungs. In the future, early diagnosis of pulmonary fibrosis should be possible. Machine learning model is helping to use the human resources efficiently and it is also reducing the expenses spent on the social and healthcare aspects of this deadly disease.

**Index Terms**—*Idiopathic Pulmonary Fibrosis (IDF), Interstitial Lung Disease (ILD), Multiple Quantile Regression, Ridge Regression, Elastic Net Regression, Machine Learning, Deep Learning, Convolutional Neural Network (CNN)*

## I. INTRODUCTION

**P**ULMONARY Pulmonary fibrosis is a common interstitial lung disease with no particular cure. It is a lung disease caused by damaged and scarred lung tissue. The lungs cannot work properly because of this thickened, stiff tissue. As pulmonary fibrosis goes to a more advance stage, it develops breathing difficulties more severe to a patient. The disease does not progress at the same rate for all patients. For some patients its progress slowly and live with PF for many years, while others it declines more quickly.

\*Correspondence Author

**Sampurna Mandal**, School of Engineering and Applied Science, The Neotia University, India Email: [sampurnamandal1564@gmail.com](mailto:sampurnamandal1564@gmail.com)

**Valentina E Balas**, Department and Applied Software, Aurel Vlaicu University of Arad, Romania, [balas@drbalas.ro](mailto:balas@drbalas.ro)

**Rabindra Nath Shaw**, Department of Electronics & Communication Engineering, Galgotias University, India Email: [r.n.s@ieee.org](mailto:r.n.s@ieee.org)

**Ankush Ghosh\***, School of Engineering and Applied Science, The Neotia University, India. Email: [ankush.ghosh@tnu.in](mailto:ankush.ghosh@tnu.in)

A patient who is suffering from pulmonary fibrosis can live from three to five years on an average after diagnosis. There are four stages of pulmonary fibrosis -- mild, moderate, severe, and very severe. Disease stage of a patient is determined after studying their lung capacity and symptoms.

However, if PF is detected in early stage and treated properly then conditions such as Pulmonary Arterial Hypertension (PAH) or Chronic Obstructive Pulmonary Disease (COPD) can impact disease prognosis.

To determine if a person is suffering an exacerbation, doctors observe his/her symptoms such as oxygen levels, CT scan results and bronchoscopy to make proper decision about diagnosis. Some of the first signs of pulmonary fibrosis are Shortness of breath, particularly during exercise, fast and shallow breathing, dry and hacking cough, tiredness, gradual unintended weight loss, clubbing (widening and rounding) of the fingertips, toes, aching joints and muscles etc.

From 1968 to 2012, 34 studies were done on 21 countries. Among which 28 studies were found to report incident data and 8 among them has revealed the mortality data. Studies show that year 2000 onwards, a 3–9 cases out 100000 cases are estimated per year for North America and Europe [1]. Most studies showed this incidence is increasing over time. The rate of IDF is growing across countries and the rate is significantly increasing worldwide. Recent data shows that the incidence shows similar conditions of liver, testicular, stomach and cervical cancers [2].

A global review has shown that the growth rate of IPF significantly lower in Asia and South America than North America and Europe. In various countries the growth rate also has differed from one region to another. Environmental or occupational risk factors can possibly be contemplated [3-7]. Studies show that the occurrence of PF is increasing significantly [1]. A current analysis based on UK-based primary care data-base between 2000 and 2012 estimated 78% rise in the incidence, with a doubling of prevalence, approximately at a rate of 38.8 out of 100,000 [6].

Studies show that IPF mortality rate is quite high. Based on survey data the patient has a high probability of 2–3 years survival from diagnosis, [8]. No improvement in survival is observed from later evidences [4,6,9]. Incidence shows that the mortality rate of IPF is increasing, even though it partly reflects increase in recognition and diagnosis [3,10-12].

## II. DATASET

When training and testing, OSIC (Open Source Imaging Consortium) Kaggle dataset is used. OSIC is a collaborative effort among academia, industry, and philanthropy. Its purpose is to confront against lung diseases by enabling rapid advances in recognition and diagnosis, including emphysematous conditions.

The csv dataset contains 2270 rows and 7 columns including Patients, Percent, Age, FVC, Sex, Weeks, and Smoking Status. This data contains 176 unique patient ids. The image folder contains 176 folders of CT scans with each file dedicated to individual patient. Each folder contains the entire CT scan history of a patient.

In the dataset, there are data for a set of patients along with their corresponding CT scan and other clinical information. The first visit of the patient is recorded as week = 0 and then is updated at the time of their next visits over the course approximately 1 to 2 year, with the measurement of FVC. In the training set, the total FVC measurements and corresponding CT scans are provided. And in the test set, only the early FVC measurements and their corresponding CT scans are provided.

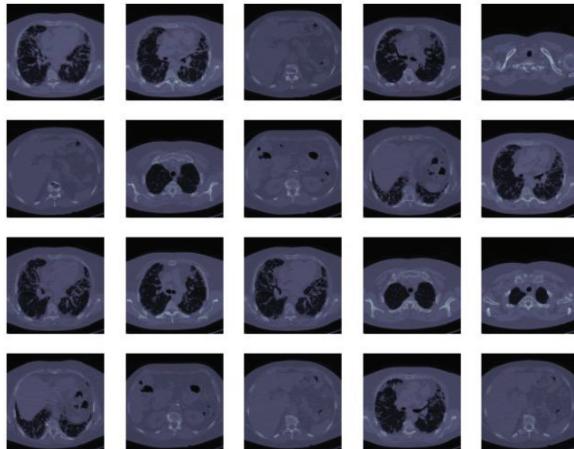


Fig. 1. Uniformity in the dataset

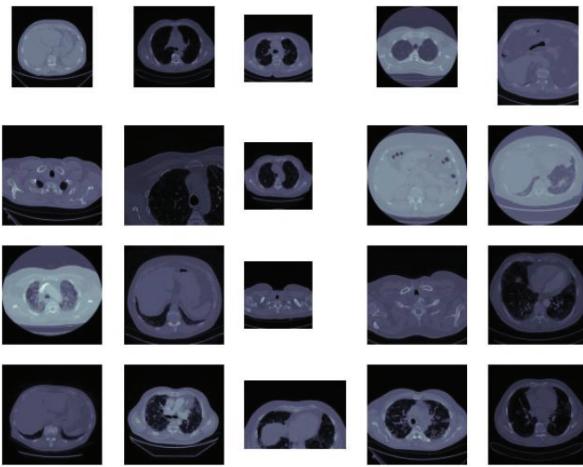


Fig. 2. Diversity in the dataset

Since it is real medical dataset, the timing of FVC measurements can vary within a wide range. Some of the CT scanned images from the dataset are depicted here:

The various training dataset characteristics are as follows:

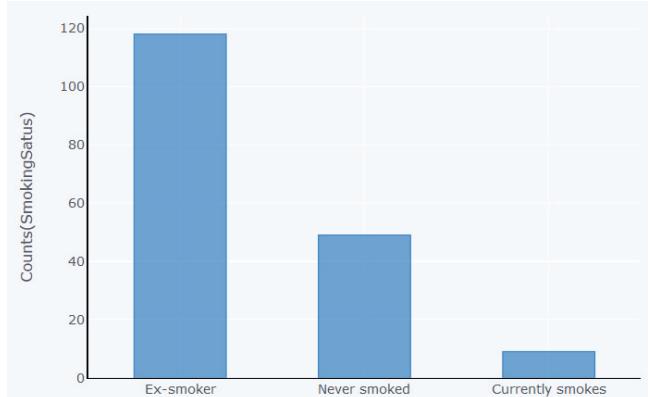


Fig. 3. Distribution of Smoking Status in the unique patient set

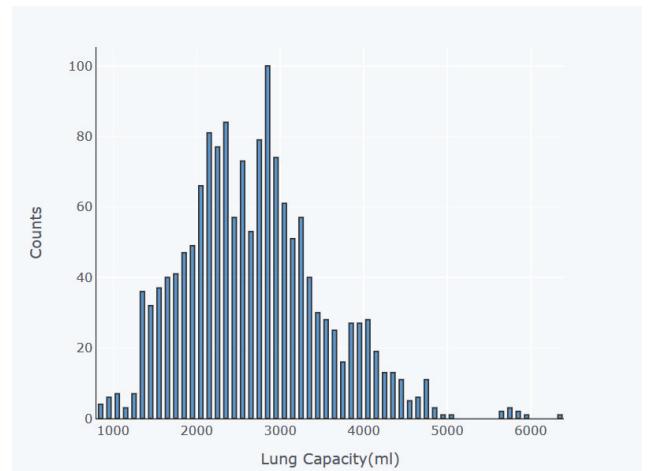
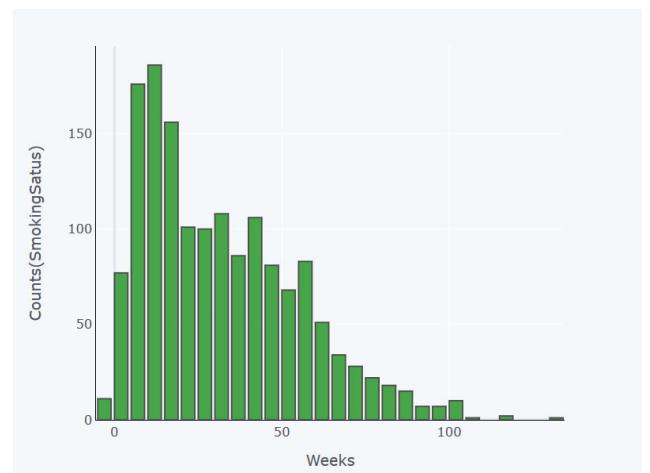


Fig. 4. a & b. Weeks vs Smoking Status in unique patient set, (b)Distribution of FVC in the training set

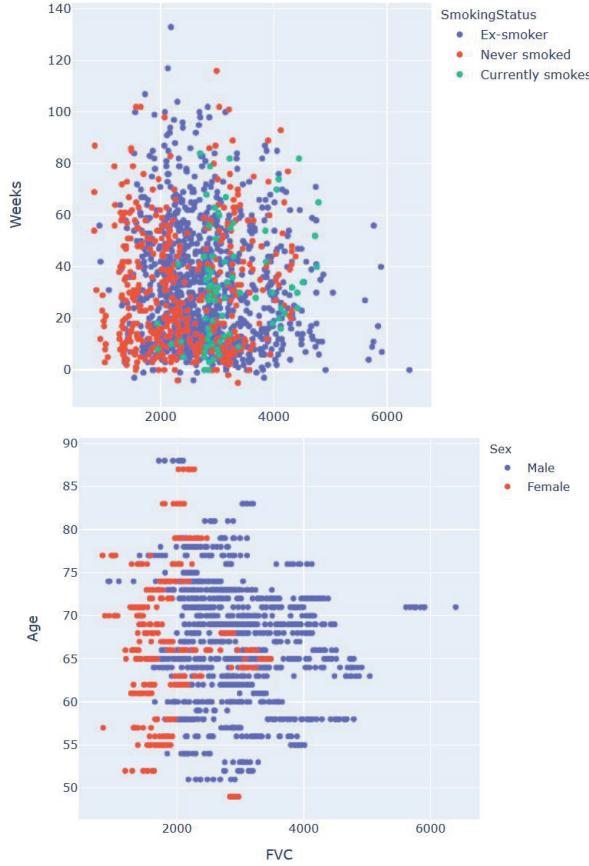


Fig. 5. (a) Distribution of FVC over age, (b) Distribution of FVC over weeks

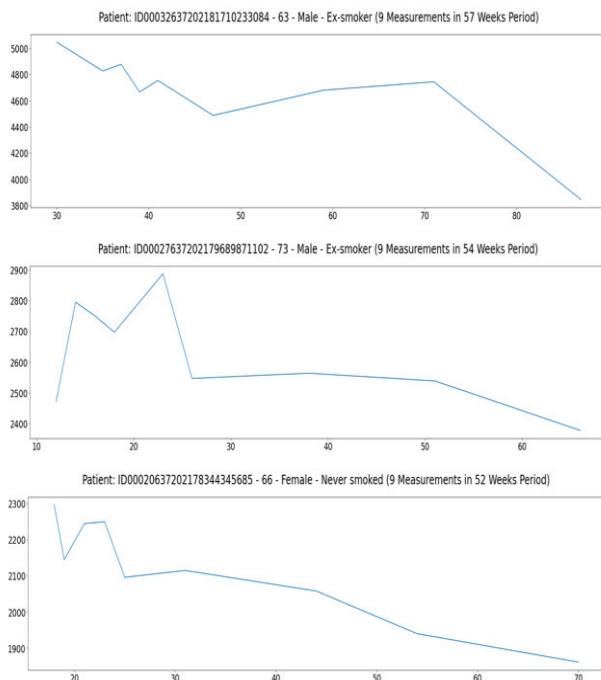


Fig. 6. FVC vs weeks graph of some patients and their corresponding age, sex and smoking status

### III. METHODOLOGY

Three models are used here – Multiple Quantile Regression, Ridge Regression and ElasticNet for training the dataset. And after that an evaluation metric is used to analyse the performance of the models. A modified version of Laplace log likelihood is used here. As Laplace log likelihood is useful evaluating the model's confidence, it is convenient for medical application. The metric returns the performance of the model in terms of accuracy and certainty for each prediction. A FVC value with a confidence (standard deviation  $\sigma$ ) is calculated for every single forced volume capacity measurement taken. The computation of the metric is given as:

$$\begin{aligned}\sigma_{clipped} &= \max(\sigma, 70) \\ \Delta &= \min(|FVC_{true} - FVC_{predicted}|, 1000) \\ metric &= -\frac{\sqrt{2}\Delta}{\sigma_{clipped}} - \ln(\sqrt{2}\sigma_{clipped})\end{aligned}$$

### IV. MODELS

#### A. Multiple Quantile Regression

Multiple quantile regression is a very beneficial statistical tool for learning the relationship between the response variable and covariates. In quantile regression, a switch is from the squared error to the tilted absolute value loss function is observed that allows to learn a specified quantile for gradient descent-based learning algorithms instead of the mean i.e. all neural network and deep learning algorithms can be applied for multiple quantile regression. In our model, Convolutional Neural Network (CNN) is used.

The equation of quantile regression is given by

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}(\alpha, \beta)} \sum_{i=1}^n \rho_\tau(y_i - f(x_i; \alpha, \beta)) + J_\lambda(f)$$

Where,  $f(x_i; \alpha, \beta) = \alpha_0 \tau + \sum_{j=1}^q \alpha_{j,\tau} z_j(x; \beta)$ ,

$\rho_\tau(v) = v(\tau - I(v < 0))$  and

$$z(x; \beta) = g^{(D)}(\cdot; \beta_D) \circ \dots \circ g^{(1)}(\cdot; \beta_1)(x) \in \mathbb{R}^q$$

For multiple linear regression, let  $0 < \tau_1 < \dots < \tau_r < 1$

The updated equation is:

$$(\widehat{f}_{\tau_1}, \dots, \widehat{f}_{\tau_r}) = \operatorname{argmin}_{f_{\tau_1}, \dots, f_{\tau_r} \in \mathcal{F}(\alpha, \beta)} \sum_{t=1}^r \sum_{i=1}^n \rho_{\tau_t}(y_i - f_{\tau_t}(x_i))$$

#### B. Ridge Regression

Tikhonov Regularization or ridge regression is a regression algorithm that is used to approximate the result (answer) for an equation which gives no unique solution. In machine learning tasks, these are one of the most common type of problem where the "best" solution must be chosen using limited data. In that case ridge regression is used for analysing multiple regression data suffering from multicollinearity. As multicollinearity arises, least squares estimates become unbiased, but their variances become large. Because of that, the calculated value may differ largely from the true value.

In Ordinary Least Square (OLS) Regression, the form of equation can be represented as follows:

$$V = \alpha_0 + \alpha_1 U_1 + \alpha_2 U_2 + \dots + \varepsilon$$

$$U_t U \alpha = U_t V$$

Where  $U$  is the design matrix having  $[U]_{ij} = U_{ij}$   $V$  is the vector containing the response ( $V_1, \dots, V_n$ ) and  $\alpha$  is the vector containing coefficients ( $\alpha_1, \dots, \alpha_n$ ).

The equation for  $\alpha$  can be written as:

$$\alpha = (U'U)^{-1}U'V$$

Where  $C = U'U$  and  $C$  is the correlation matrix of independent variables.

Ridge regression adds a small value  $\varphi$ , to the diagonal elements of  $C$  which is basically correlation matrix.

$$\alpha = (C + \varphi I)^{-1}U'V$$

Where the value of  $\varphi$  lies in the range of  $0 < \varphi < 1$ .

### C. ElasticNet Regression

ElasticNet is a linear model which is used in Machine Learning. ElasticNet regularization implements both L1-normalization and L2-normalization regularization to penalize the coefficients in a regression model.

$$\hat{\alpha}^{ridge} = \operatorname{argmin} \|m - Y\alpha\|_2^2 + \lambda\|\alpha\|_2^2$$

$$\hat{\alpha}^{lasso} = \operatorname{argmin} \|m - Y\alpha\|_2^2 + \lambda\|\alpha\|_1$$

$$\hat{\alpha}^{elastic} = \operatorname{argmin} \|m - Y\alpha\|_2^2 + \lambda_2\|\alpha\|_2^2 + \lambda_1\|\alpha\|_1$$

Both Ridge and Elastic Net Regression belong to the same family including penalty term of:

$$P_\beta = \sum_{i=1}^p \left[ \frac{1}{2} (1-\beta) b_j^2 + \beta |b_j| \right]$$

For Ridge Regression  $\beta = 0$  and for Elastic Net Regression  $0 < \beta < 1$ .

## V. RESULTS AND DISCUSSIONS

### A. Train Test Plot



Fig. 7. Train Test plot for (a) Multiple Quantile Regression (b) Ridge regression and (c) Elastic Net Regression

In figure 7, the blue and the orange curves represents train and test curve respectively. It shows how well the training and testing curve fits.

### B. FVC Distribution and OOF Confidence Distribution

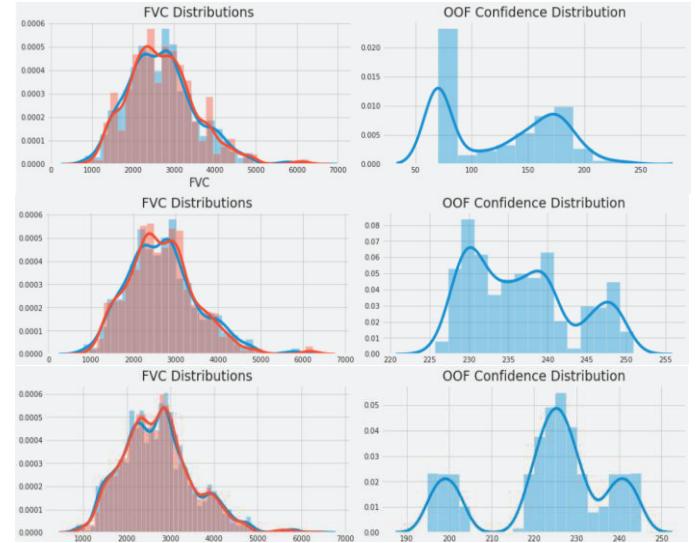


Fig. 8. FVC Distribution and OOF Confidence Distribution for(a) Multiple Quantile Regression (b) Ridge regression and (c) Elastic Net Regression

In Fig. 8, in FVC Distribution figures the blue and the orange curves and bars represents train and test curve respectively. OOF Confidence Distribution curves show the

confidence of the model according to the OOF score of the corresponding model.

The OOF (out-of-fold) score for (a) Multiple Quantile Regression (b) Ridge regression and (c) Elastic Net Regression models are shown below:

### C. OOF (out-of-fold) score for various models

TABLE I. OOF SCORE

Model name	OOF SCORE
Multiple Quantile Regression	-6.92
Ridge Regression	-6.81
Elastic Net Regression	-6.73

## VI. CONCLUSIONS

In this paper we have shown the comparison among various ML (Machine learning) models' performance to analyse Pulmonary Progression. This type of prediction analysis, if applied effectively in medical sectors, can help patients by analysing their lungs condition from Ct scan and the corresponding other information so that treatment can be started as early as possible. Early treatment can improve the survival rate of a patient. Thus, ML can help medical practitioners to understand their prognosis in a better way when they are first diagnosed with IPF. Hence, machine learning algorithms are adding significant value to the healthcare industry as well as to the society aiming towards a healthy and normal social lifestyle.

## REFERENCES

- [1] Hutchinson J., Fogarty A., Hubbard R., McKeever T. Global incidence and mortality of idiopathic pulmonary fibrosis: A systematic review. *Eur. Respir. J.* 2015;46:795–806. doi: 10.1183/09031936.00185114.
- [2] Stephen M. Humphries, Jeffrey J. Swigris, Kevin K. Brown, Matthew Strand, Qi Gong, John S. Sundy, Ganesh Raghu, Marvin I. Schwarz, Kevin R. Flaherty, Rohit Sood, Thomas G. O'Riordan, David A. Lynch. Quantitative high-resolution computed tomography fibrosis score: performance characteristics in idiopathic pulmonary fibrosis. *European Respiratory Journal* 2018; 52: 1801384; DOI: 10.1183/13993003.01384-2018
- [3] Fleming K.M., Navaratnam V., West J., Smith C.J., Hubbard R.B, Jenkins R.G., Fogarty A., The rising incidence of idiopathic pulmonary fibrosis in the U.K. *Thorax*. 2011;66:462–467. doi: 10.1136/thx.2010.148031.
- [4] Yeh W.S., Lee Y.C., Raghu G., Chen S.Y., Maroni B., Li Q., Collard H.R. Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: Incidence, prevalence, and survival, 2001–11. *Lancet Respir. Med.* 2014;2:566–572. doi: 10.1016/S2213-2600(14)70101-8.
- [5] Hopkins Burke N., Dion G.R.B., Kolb M. Fell C. Epidemiology and survival of idiopathic pulmonary fibrosis from national data in Canada. *Eur. Respir. J.* 2016;48:187–195. doi: 10.1183/13993003.01504-2015.
- [6] Maher T.M Strongman H., Kausar I. Incidence, Prevalence, and Survival of Patients with Idiopathic Pulmonary Fibrosis in the UK. *Adv. Ther.* 2018;35:724–736. doi: 10.1007/s12325-018-0693-1.
- [7] Gribbin J., Hubbard R.B., le Jeune I., Smith C.J., West J., Tata L.J. Incidence and mortality of idiopathic pulmonary fibrosis and sarcoidosis in the UK. *Thorax*. 2006;61:980–985. doi: 10.1136/thx.2006.062836.
- [8] Hutchinson J., Fogarty A., Hubbard R., McKeever T. Global incidence and mortality of idiopathic pulmonary fibrosis: A systematic review. *Eur. Respir. J.* 2015;46:795–806. doi: 10.1183/09031936.00185114.
- [9] Kelloniemi K., Kaunisto J., Sutinen E., Hodgson U., Piilonen A., Kaarteenaho R., Makitalo R., Purokivi M., Lappi-Blanco E., Saarelainen S., et al. Re-evaluation of diagnostic parameters is crucial for obtaining accurate data on idiopathic pulmonary fibrosis. *BMC Pulm. Med.* 2015;15:92. doi: 10.1186/s12890-015-0074-3.
- [10] Saliccioli J.D., Marshall D.C., Akuthota P., Shea B.S. Trends in mortality from idiopathic pulmonary fibrosis in the European Union: An observational study of the WHO mortality database from 2001–2013. *Eur. Respir. J.* 2018;51 doi: 10.1183/13993003.01603-2017.
- [11] Algranti E., Saito C.A., Silva D., Carneiro A.P.S., Bussacos M.A. Mortality from idiopathic pulmonary fibrosis: A temporal trend analysis in Brazil, 1979–2014. *J. Bras. Pneumol.* 2017;43:445–450. doi: 10.1590/s1806-37562017000000035.
- [12] Diamantopoulos A., Wright E., Vlahopoulou K., Cornic L., Schoof N., Maher T.M. The Burden of Illness of Idiopathic Pulmonary Fibrosis: A Comprehensive Evidence Review. *Pharmacoeconomics*. 2018 doi: 10.1007/s40273-018-0631-8.
- [13] Naga Srinivasu, P., Balas, V.E., Md. Norwawi, N., “Performance measurement of various hybridized kernels for noise normalization and enhancement in high-resolution MR images”, *Studies in Computational Intelligence* 2021
- [14] Balas, M.M., Balas, V.E., Lile, R., Balas, S.V., “Fuzzy-Interpolative Control for Intelligent Roof-Top Greenhouse Buildings” *Studies in Fuzziness and Soft Computing*, 2021
- [15] Milan Kumar, V. M. Shenbagaraman and Ankush Ghosh, “Innovations in Electrical and Electronic Engineering” Book Chapter, Springer. [ISBN 978-981-15-4691-4, Favorskaya et al (Eds.): Innovations in Electrical...]