Accurate Prediction of Pulmonary Fibrosis Progression Using EfficientNet and Quantile Regression: A High Performing Approach

Rofiqul Alam Shehab, Kaysarul Anas Apurba, Md. Ahsanuzzaman, Dr. Tanzilur Rahman, IEEE

***Abstract*— Pulmonary fibrosis (PF) is a chronic lung disease characterized by the formation of scar tissue in the lungs, leading to difficulty breathing and a reduced ability to oxygenate the blood. The progression of PF is difficult to predict, and current methods of diagnosis and treatment are often ineffective. In this study, we propose to use EfficientNet, utilizing a cutting-edge convolutional neural network (CNN) architecture and quantile regression (QR) to predict the progression of PF in patients. Our approach includes analyzing data from the OSIC dataset, the biggest publicly accessible dataset containing medical imaging, patient demographics, and lab results. The analyzed data was trained on an EfficientNet model and QR to predict the progression of the disease, as well as estimate the uncertainty of the predictions. The performance of the model was evaluated using Laplace-Log-Likelihood. The results demonstrate that the proposed approach outperforms existing literature in predicting pulmonary fibrosis progression, with the highest score (-6.64). This approach has the potential to aid in the development of new treatments for this disease**.

***Keywords*—IPF, pulmonary fibrosis, neural network, computed tomography, chest CT, convolutional, pertained models, open source.**

Authors are all with the North South University. Bashundhara, Dhaka-1229, Bangladesh (corresponding author to provide phone: +880-2-55668200; e-mail: tanzilur.rahman@northsouth.edu).

I. INTRODUCTION

Pulmonary fibrosis is a chronic lung disease characterized by the formation of scar tissue in the lungs, resulting in difficulty breathing and a reduced ability to oxygenate the blood. Predicting the progression of pulmonary fibrosis (PF) is critical for developing new treatments and managing patients with the disease. Evaluating lung function decline is critical in treating and managing pulmonary fibrosis. Most methods for diagnosing pulmonary fibrosis are based on guidelines published in an official ATS/ERS/JRS/ALAT clinical practice guideline journal [1], such as transbronchial lung biopsy, surgical lung biopsy, spirometry tests, etc. The transbronchial [2] and surgical biopsy [3] methods are used to diagnose pulmonary fibrosis, and there are risks to the patient's health and lung function. Spirometry methods are well suited to assessing the decline in lung function following a diagnosis because they are widely used to quantify the lung's forced vital capacity (FVC), a critical marker of lung function [4]. However, with only the spirometry test, it is difficult to understand the disease state, severely limiting its ability to predict the development of pulmonary fibrosis. As medical technology has advanced over the last few decades, it has become more useful in clinical decisions. Developing novel techniques for improving prediction precision is highly desirable when using CT images to evaluate and forecast future lung function decline caused by pulmonary fibrosis. Machine learning and deep learning can help in this situation, thanks to recent advancements and the possibility of artificial intelligence techniques in the medical sector. Existing PF diagnosis and treatment methods are frequently ineffective, and disease progression is difficult to predict. In this study, EfficientNet, a cutting-edge convolutional neural network (CNN) architecture, and quantile regression (QR) are proposed to predict the progression of PF in patients using the OSIC dataset [5]. Our approach includes the following key contributions:

- Using the OSIC dataset, including medical imaging, patient demographics, and lab results, to forecast disease progression.
- Combining EfficientNet and Quantile Regression to achieve a more accurate and robust prediction of PF progression, as well as an estimate of the prediction's uncertainty.
- Evaluating our models' performance with the modified Laplace-Log-Likelihood metric, which is better suited for predicting continuous outcomes with censoring.
- Our model's best score in terms of modified Laplace-Log-Likelihood (-6.64) is higher than the existing literature.

Our approach has the potential to aid in the development of new treatments for this disease as well as the management of PF patients.

II. RELATED WORK

Yu Shi et al. were the first to demonstrate that artificial intelligence could predict the progression of idiopathic pulmonary fibrosis using only baseline HRCT scans.

2023 IEEE Region 10 Symposium (TENSYMP) | 978-1-6654-8258-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/TENSYMP55890.2023.10223673

Using only baseline HRCT scans, they attempted to develop a novel predictive model for the radiological progression pattern of idiopathic pulmonary fibrosis. Their proposed model has an overall accuracy rate of 82.1 percent, which is higher than the other feature selection and classification methods discussed above. The Kaggle Pulmonary Fibrosis Progression Challenge [5] recently demonstrated the potential and need for breakthroughs in pulmonary fibrosis artificial intelligence-driven solutions and computer-aided clinical decision support. This challenge was issued by the Open-Source Imaging Consortium (OSIC) [6] to encourage the scientific community to accelerate the development of machine learning for pulmonary fibrosis evaluation. The winning solution [7] proposed a weighted ensemble of a deep convolutional neural network with a cutting-edge EfficientNet-B5 network architecture design and multiple quantile regression. Wong et al. [8] created a tailored network design for predicting forced vital capacity (FVC). They carried out their experiment using a CT scan, an initial spirometry measurement, and clinical metadata from a patient. They identified a robust architectural design for CT lung analysis using machine-driven design exploration. Finally, they use an explainability-driven performance validation technique to investigate the fibrosis-decision-making network's behavior and confirm that predictions are based on relevant visual cues in CT images. They received the highest score from all previous experiments during their investigation. The Fibro-CoSANet [9] model employs convolutional layers as well as self-attention mechanisms. Convolutional layers are widely used in image analysis and can identify important features in medical images such as CT scans. The self-attention mechanism enables the model to weigh various factors and identify important patterns in the data, potentially resulting in more accurate prognosis predictions. The Fibro-CoSANet model represents a significant advancement in predicting the prognosis of pulmonary fibrosis. The model's use of convolutional and self-attention layers enables it to analyze patient data and make accurate predictions about disease progression. However, more research is needed to validate these findings and test the model in larger and more diverse patient populations in order to improve the model's generalizability. Despite the fact that a substantial amount of work has been completed, extensive research has revealed that there is still significant room for improvement. As a result, the goal was to provide a new and improved system that outperformed previous versions. It is a system that could be extremely beneficial in encouraging more open and consistent clinical adoption of those solutions.

III. Methodology

A. Dataset: Benchmark Dataset for the OSIC Pulmonary Fibrosis Progression Challenge includes CT scans, demographic information, and FVC measurements such as gender, age, and smoking status. Based on a patient's CT scan of their lungs, metadata, and baseline FVC as input, the dataset has 2270 rows and seven columns. There are approximately 200 cases in the combined public and private test sets. This data set includes 176 unique patient identifiers. There are 176 CT scan folders in the image folder, with each file dedicated to a specific patient. Each folder contains the entire CT scan history of a patient. Our goal is to predict how severely their lung function will deteriorate. Aim to forecast the last three FVC measures and a prediction confidence level for each patient.

B. Preprocessing: With a few minor outliers, the data is divided into two size groups. CT scans with 512x512-pixel images were performed on 132 patients. There are 34 patients in total at this size, with the remaining being 1 or 2. Because the largest piece of data is that size and contains 34000 images, each one must be resized to 512 x 512. Then I attempted to draw conclusions based on three characteristics: age, gender, and smoking status. The CT scan file must be converted to Hounsfield units because the x-ray spectrum is determined by measurement settings such as acquisition parameters and tube voltage. By normalizing the values of water and air, the images of various measures can be compared (water has HU 0 and air -1000). About 4,000 of the gray values a CT scan produces are invisible to our eyes. This is the reasoning behind windowing. As a result, the image is displayed in the HU range that best suits the area of interest. A voxel is a 3D pixel generated by a CT scan. According to our understanding, the slice thickness spans it in the z-direction and the pixel spacing property's 2D plane in the x- and y-directions. The pixel spacing characteristic of dicom files is critical. It reveals the actual distance covered by one pixel. A transversal slice's x- and y-axes are described by only two values in the plane. This pixel spacing is frequently consistent across all slices of a single patient. However, the pixel spacing between patients can vary depending on the individual or organizational preferences of doctors and clinics and the type of scanner used. As a result, even when two photographs of the same organ size are compared, the larger image does not always reflect the actual size of the organ. By measuring the thickness of a slice, one can determine how far it travels in the z-direction. The pixel array of raw values also covers a specific area defined by row and column values. In very thin slices, more features

can be seen. Thick slices, on the other hand, have less noise but are prone to artifacts.

C. System Architecture: The proposed method is divided into two parts. Create a model first that combines categorical data with image data. In the second part, the system uses the model to generate a new data frame. Create the final model to forecast the progression using that data set after some adjustments and calculations.

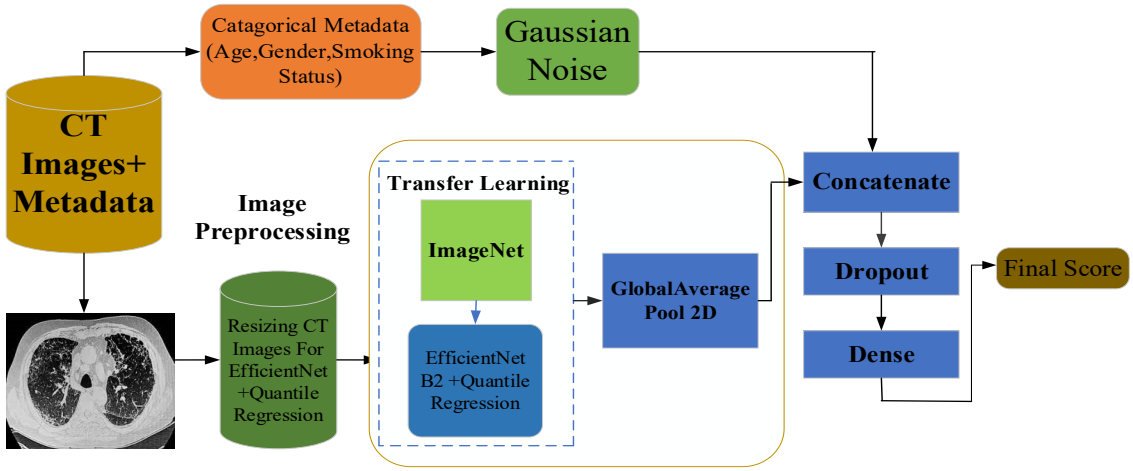

Figure 1. The Proposed System architecture for prediction of FVC.

Given a patient's CT scan and metadata (age, gender, smoking status), our system "Fig. 1" predicts the patient's forced vital capacity at a specific week in the future.

In the first step, this system collects and prepares every image for training. The image data must be trained quickly to avoid running out of memory. Create an image generator to train a single piece of data at a time while avoiding errors. The model is then constructed. The model's architecture is shown in "Fig. 1." Because the input layers of all models are 224*224. After that, create a model and add an input layer with the dimensions 512*512*1. The pre-trained model is then fed with image data, and a global average pooling 2D layer is added. Age, gender, and smoking status are categorical variables processed through a Gaussian noise layer before being analyzed. [10] These tasks were completed concurrently. Combine the two, then add a dropout layer and pass the result through a dense layer. Following that is the training phase. Early stopping is used during the training phase. We use it because using too many epochs can result in an overfit model, whereas using too few can result in an underfit model. Early stopping is a technique that allows us to specify an infinite number of training epochs and terminate training when the model's performance on a holdout validation dataset stops improving. When the metrics improve, we lower the learning rate on the plateau. Models typically gain by slowing down learning once the learning rate reaches a plateau during the learning process. This callback keeps track of a quantity, and if no progress is made after 20 epochs, the learning rate is reduced.

Table 1 Example of Submission Table OSIC Challenge

Patient_week	FVC	Confidence
ID00419637202311204720264 -12	3111.484416	161.671271
ID00421637202311550012437 -12	2857.890357	200.935648
ID00422637202311677017371 -12	2007.297131	153.969624
ID00423637202312137826377 -12	3420.942369	206.201271
ID00426637202313170790466 -12	2974.510875	121.335843

In the second part, the model generated two variables, tr for train and chunk for test, for storing train and test sets. It included the sample submission data in a variable called "sub" because it tends to forecast the FVC value by week. (Table I) shows three characteristics: "patient week," "FVC," and "confidence." As can be seen, each entry in the "patient week" field is in concatenated form. To save them, we separated them into "patient id" and "week" fields and added new features to the "sub" table. We added a "where" feature to the "tr," "chunk," and "sub" tables and initialised it with "train," "val," and "test," respectively. Then combine them all into a single table called "data." We added a new feature called "start week" to the "data" table and assigned "nan" values (which stand for "not a number") to "start week" in accordance with the "test" rows.

Table II Showing each patient's opening week FVC values.

Patient_I	Start_FVC
ID00007637202177411956430	2315
ID00009637202177434476278	3660
ID00010637202177584971671	3523
ID00011637202177653955184	3326

Following that, we determined the arithmetic mean of the weeks for each patient and assigned a "start week" column to each patient based on that value. Then, to obtain the starting week FVC value, it establishes a new table named "base" which only includes the "start week" for each patient. This table is depicted in (Table II). The "base" table contains the initial weeks' fvc value for each table. In terms of patient ID, then combine them with the "data" table. The data table now has a new feature called "base week."

Table III Displaying the Patient's Gender Based Smoking Status

Patient_ID	Male	Female	Ex-smoker	Never smoked	Currently smokes
ID00419637202311204720264	1	0	1	0	0
ID00421637202311550012437	1	0	0	0	1
ID00422637202311677017371	0	1	0	1	0
ID00423637202312137826377	1	0	0	1	0

We subtract "week" from "start week" for each entry and enter the result in "base week." Because the system recognized that "sex" and "smoking status" are categorical, I created those characteristics in the table and converted them to Boolean . As shown in (Table III), the corresponding entry is 1 if the patient is male and 0 otherwise, and the same is true for smoking status. For each patient, the values for "age," "fvc," "week," and "percent" have now been normalized and added to the "feature" table. To simplify the calculation, I ran all of these analyses in a single data frame before dividing them into train, validation, and test sets. Then, in order to train the new dataset shown in (Table II), we needed to create a new model. The model's categorical and numerical data were processed and passed through several dense layers. [11] Aside from that, we used one hot encoding to transform the category values to Boolean values. The model was then trained and validation data was used to confirm its accuracy. The section that follows goes into detail about training, validation, and testing.

Evaluation Metric: Laplace Log Likelihood

To assess our project, a modified version of the Laplace log likelihood will be used. It is helpful to assess a model's confidence in its conclusions when used in medical applications. As a result, the measure is made to reflect both each prediction's accuracy and certainty. For each true FVC measurement, predict both an FVC and a confidence measure (which is standard deviation σ). The metric is determined by:

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

$$\Delta = \min(|FVC_{true} - FVC_{predicted}|, 1000),$$

$$\text{Metric} = -\frac{\sqrt{2}\Delta}{\sigma clipped} - \ln(\sqrt{2}\sigma_{clipped})$$

IV. Result Analysis

EfficientNet (B0, B1, B2, and B4) with quantile regession, ResNet50, and VGG16 were trained for experimentation. Each of these models has different parameters and architecture. These models were trained with the same set of hyperparameters to allow for comparison.The lowest training loss for EfficientNet B0 with Quantile regression is 3.6524. The highest training loss for ResNet50 is 4.399539, whereas the training loss for EfficientNet B1' with quantile regression is 4.2689. As shown in, the training loss for all six models is nearly identical (Table IV). However, the training loss for ResNet50 and VGG16 is slightly lower.

When testing the model, however, it is discovered that ResNet50 and VGG16 need to be more balanced. Because our dataset is small and ResNet50 and VGG16 have complex architectures that require a large amount of training data, there is a chance that the data will overfit. The training accuracy for each of the seven models over a 100-epoch period is shown in (Table IV). When looking at each graph in "Fig. 2," it cannot see any variations in training loss. This is because the difference in training losses is so small. The training and validation losses of these models on smooth curves were then illustrated in "Fig. 3," respectively, to help alleviate the problem.

The performance of our trained model was then evaluated using a modified version of the Laplace-Log-Likelihood. When used in medical applications, assessing a model's confidence in its conclusions is useful. The confidence values are clipped at 70 ml to reflect the approximative measurement uncertainty in FVC, and the error is thresholded at 1000 ml to prevent large errors from negatively influencing results. The final score is calculated by averaging the statistic overall test set's patient weeks (three per patient). Remember that metrics will have negative values, with higher being better. The Laplace-log- likelihood function was created to comprehend the metric and apply it to various cases. The default score to beat is -8.023 when cross-validating models with train data. Any model that performs worse than this is rendered ineffective.

Table IV Comparison of Training Loss with Different Model

Model Name	Training Loss	Number of epochs
EfficientNet B0+Quantile Regression	3.6524	100
EfficientNet B1+Quantile Regression	4.2689	150
EfficientNet B2+Quantile Regression	3.8340	100
EfficientNet B3+Quantile Regression	3.6728	100
EfficientNet B4+Quantile Regression	4.1240	100
Res Net 50	4.3995	100
VGG16	4.1878	100

Except B1, the model has been trained over 100 epochs and has undergone five-fold cross-validation. The Kaggle competition's scoring system was used for validation. Initially, they only looked at the previous three forecasts (the previous three weeks for each patient). As a result, a comparable validation framework was created. The test set included only patients who did not participate in the training set, and only the three most recent weeks were graded. In the evaluation, the highest possible score was -6.6468.
(Table V) compares our results to those in the study as well as the public Kaggle leaderboard. The comparison table has been arranged in ascending order to determine the highest score. As we can see, it outperformed the competition and previously published work.

Table V Represent the comparison between our score of with all Existing Scores

Model Name	Scores
Our Model (EfficientNet B2 + Quantile Regression)	-6.64
Fibro-CoSANet [9]	-6.68
Elastic Net Regression [12]	-6.73
Fibrosis-Net [8]	-6.8188
Kaggle 1st place [7]	-6.81
Kaggle 2nd place [5]	-6.83
Kaggle 3rd place [5]	-6.83
Ridge Regression [12]	-6.81
Multiple Quantile Regression [12]	-6.92

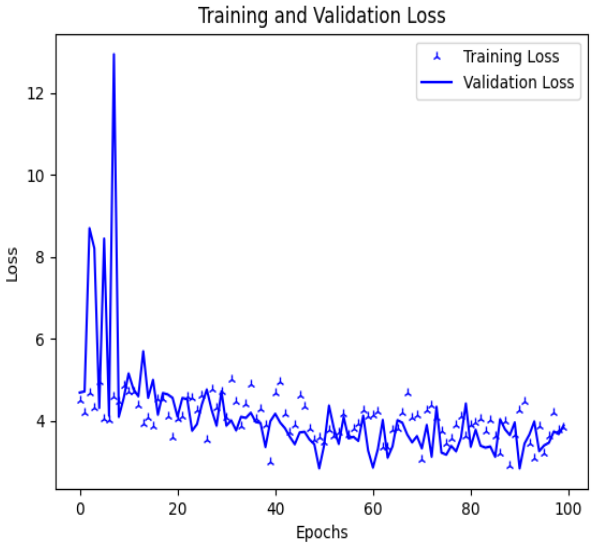

Figure 2- Showing the Training and validation loss of EfficientNet B2 and Quantile Regression

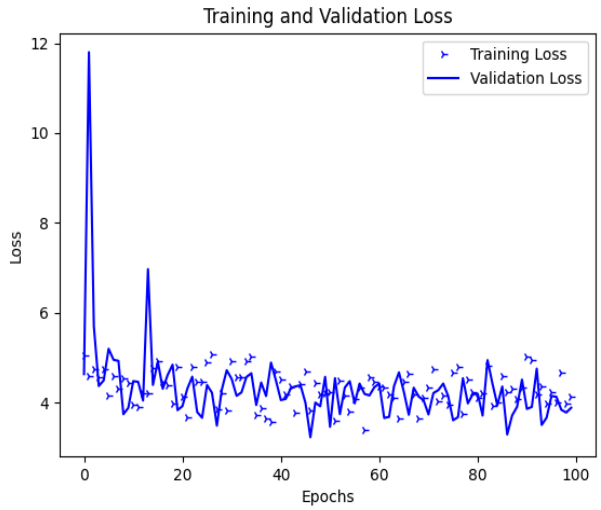

Figure 3- Showing the Training and validation loss of EfficientNet B4 and Quantile Regression

V. Conclusion

Our proposed work used metadata and CT-scan images to predict FVC and measured its performance using a Laplace-Log-Likelihood evaluation matrix. Our achieved score is -6.64, which outperforms all other existing models. Based on our findings, high-resolution CT can be combined with our proposed deep learning techniques to provide a low-cost, efficient, fast, and appropriate method of detecting a decline in lung function in a patient with idiopathic pulmonary fibrosis. This will allow not only the patient, but also the clinical specialist, to make better decisions more quickly.

VI. References

[1] G. R.-J. M. M. J. L. R. L. R. C. J. L. D. J. Raghu, "Diagnosis of Idiopathic Pulmonary Fibrosis. An Official Ats/ers/jrs/alat Clinical Practice Guideline," *Am. J. Respir. Crit. Care Med,* no. doi:10.1164/rccm.201807-1255st, pp. 198, e44–e68., 2018.

[2] L. C. H. R. a. J. M. G. Richeldi, "Idiopathic Pulmonary Fibrosis.," *The Lancet,* no. doi:10.1016/s0140-6736(17)30866-8, pp. 389, 1941–1952, 2017.

[3] S. W. A. U. C. U. C. A. C. T. V. R. G. Tomassetti, "Bronchoscopic Lung Cryobiopsy Increases Diagnostic Confidence in the Multidisciplinary Diagnosis of Idiopathic Pulmonary Fibrosis," *Am. J. Respir. Crit. Care Med.,* no. doi:10.1164/rccm.201504-0711oc, pp. 193, 745–752, 2016.

[4] L. C. K. T. E. S. M. I. W. J. A. S. R. E. a. C. R. M. Watters, "A Clinical, Radiographic, and Physiologic Scoring System for the Longitudinal Assessment of Patients with Idiopathic Pulmonary Fibrosis.," *Am. Rev. Respir. Dis,* no. doi:10.1164/arrd.1986.133.1.97, pp. 133, 97–103, 1985.

[5] O. S. I. C. (OSIC), "OSIC Pulmonary Fibrosis Progression-Predict lung function decline," 2019. [Online]. Available: https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression.

[6] OSIC, "Open Source Imaging Consortium (OSIC)," [Online]. Available: https://www.osicild.org/.

[7] A. Kulakov, "How I achieved the 1st place in Kaggle OSIC Pulmonary Fibrosis Progression competition," Medium, 7 Oct 2020. [Online]. Available: https://towardsdatascience.com/how-i-achieved-the-1st-place-in-kaggle-osic-pulmonary-fibrosis-progression-competition-e410962c4edc. [Accessed 19 January 2022].

[8] J. L. A. D. P. M. M. F. D. M. J. R. H. Alexander Wong, "Fibrosis-Net: A Tailored Deep Convolutional Neural Network Design for Prediction," *Frontiers in Artificial Intelligence,* vol. 4, no. DOI=10.3389/frai.2021.764047, 2021.

[9] F. R. M. ,. M. A. I. ,. S. S. Zabir Al Nazi, "Fibro-CoSANet: Pulmonary Fibrosis Prognosis Prediction using a Convolutional Self Attention Network," *Physics and Engineering in Medicine,* no. DOI: 10.1088/1361-6560/ac36a2, 2021.

[10] J. Brownlee, "How to Improve Deep Learning Model Robustness by Adding Noise," Machine Learning Mastery, 14 December 2018. [Online]. Available: https://machinelearningmastery.com/how-to-improve-deep-learning-model-robustness-by-adding-noise/. [Accessed 20 january 2023].

[11] D. Kaplan, "Dense Layer: The Building Block To Neural Networks," EML, 7 June 2022. [Online]. Available: https://enjoymachinelearning.com/blog/dense-layer/. [Accessed 20 January 2023].

[12] S. Mandal, V. E. Balas, R. N. Shaw and A. Ghosh, "Prediction Analysis of Idiopathic Pulmonary Fibrosis Progression from OSIC Dataset," in *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON),*, Greater Noida, India, 2020.