Prediction Of Pulmonary Fibrosis Disease

Disha Jain
Dept. of Computer Science and Engineering
Amity University
Uttar Pradesh, India
jaindisha2101@gmail.com

Palak Khurana
Dept. of Computer Science and Engineering
Amity University
Uttar Pradesh, India
palakkhurana06@gmail.com

Sakshi Yadav
Dept. of Computer Science and Engineering
Amity University
Uttar Pradesh, India
yanya492@gmail.com

Seema Sharma
Dept. of Computer Science and Engineering
Amity University
Uttar Pradesh, India
ssharma26@amity.edu

2024 Control Instrumentation System Conference (CISCON) | 979-8-3503-7548-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/CISCON62171.2024.10696824

***Abstract*-The progressive and potentially fatal lung illness pulmonary fibrosis (PF) is represented by permanent scarring in a lung tissue. A successful treatment of pulmonary function decline requires early diagnosis and close observation. For people with Pulmonary fibrosis (PF), forced vital capacity (FVC), a measurement of maximal exhaled air volume, is an essential marker of lung function. In order to improve PF diagnosis, this work investigates the viability of using deep learning and machine learning models to predict FVC. A set of three models were developed namely the Long Short Term-Memory (LSTM) network to identify sequential patterns in data from clinical trials, a Huber regression model for reliable management of outliers, and a Support Vector Regression (SVR). For model training and assessment, a dataset of 200 patients was used. In the training set, baseline Computed Tomography (CT) scan and the entire history of FVC measurements are provided and, in the testing, set baseline CT scan and only the initial FVC measurements are used. Each model's effectiveness was compared using three performance metrics- Mean Absolute Error, Mean Squared Error and Root Mean Squared Error. The findings showed that all three of the models had potential for predicting FVC, with the SVR be the most efficient in predicting the FVC values as it has the lowest Mean squared error (MSE) which is 0.029. This work demonstrates the potential of deep learning and machine learning for non-invasive FVC prediction, which may allow for early PF detection and prompt intervention for improved clinical results.**

Keywords—Capacity, Computed Tomography, Forced Value Capacity, Hubber regression, Long short term-memory, Mean squared error, pulmonary fibrosis, Support Vector Regression

I. INTRODUCTION

A chronic and damaging lung illness called pulmonary fibrosis (PF) is defined by the lungs' gradual buildup of scar tissue. Because of this scarring, the transfer of gases is obstructed, which eventually results in respiratory failure and a reduced ability to tolerate exercise. Due to its high mortality rate and restricted treatment choices, Pulmonary Fibrosis poses a considerable healthcare burden. Enhancing patient outcomes and quality of life requires prompt detection and treatment.

Clinical symptoms, high-resolution computed tomography (HRCT) scans, and invasive surgical lung biopsies have historically been used to diagnose the disease. However, lung biopsies come with dangers and problems, and HRCT scans are vulnerable to subjectivity and inter-reader variability. Thus, non-invasive, objective techniques are desperately needed to help in early PF identification.

One reliable metric of lung function is Forced Vital Capacity (FVC), which is the greatest volume of air that an individual can forcibly expel after taking a deep breath. More severe PF is linked to lower FVC readings. Accurately predicting FVC may be a useful tool for clinicians in diagnosing and tracking the evolution of Pulmonary Fibrosis disease.

The paper will compare the performance of a Long Short-Term Memory (LSTM) network for identifying sequential patterns in clinical data, a Huber regression model for reliable handling of outliers, and a Support Vector Regression (SVR) for regression. It has been observed that these models are capable of predicting nearly accurately FVC values, which could lead to earlier and more successful diagnosis and treatment of the disease.

II. BACKGROUND STUDY

This section focuses on thorough review of the study on the Pulmonary Fibrosis prediction using machine learning and deep learning methods. A study [1] describes how deep neural networks can be used to predict lung fibrosis using X- rays. It talks about how difficult it is to diagnose pulmonary fibrosis. It also describes a system that classifies the segmented lung tissue using a different neural network and segments lung tissue in X-rays using a neural network. The authors suggest a two-phase approach where lung segmentation is used to separate the lung region from the chest X-ray, a deep learning model is used in the first stage. Precise segmentation is essential for later examination of lung tissue properties.PF classification is done to help with PF diagnosis, the second step uses a different deep learning model to categorize the segmented lung tissue as either PF positive or negative.

Another study [2] provides significant information about automated systems. By automating PF categorization and lung segmentation, the suggested system may shorten analytical times and increase workflow effectiveness. However, the model's efficacy is acknowledged by the authors to be inferior to that of licensed radiologists. The research suggests two methods for examining the development of IPF: Model 1: Histogram features taken from HRCT scans are combined with biological information about the patients. An elementary statistical examination of the image's intensity distribution is offered by histogram characteristics.

Model 2: Using an ensemble method, this model combines quantile regression on biological data from patients with a Convolutional Neural Network (CNN) trained on HRCT scans. CNNs are an effective deep learning architecture that can identify intricate patterns in photos.

The results of the study indicate that Model 2, which uses an ensemble approach to combine biological and imaging data, obtains a stronger correlation with FVC values than Model 1. This suggests that Model 2 might provide a more reliable and insightful method for forecasting the course of IPF. Furthermore, the research indicates that biological information can be more important than HRCT scans for estimating FVC reduction in PF patients.

A paper by Alwani and others [3] focuses on the development of a machine learning model for early PF detection. The essential

979-8-3503-5098-2/24/$31.00 ©2024 IEEE

elements consist of the model selection to determine which machine learning algorithm has the best prediction accuracy for IPF, researchers will examine a number of options, such as Random Forest, Support Vector Machine (SVM), Naive Bayes, and J48. creating and evaluating a two-layered deep neural network (DNN) model for binary classification (IPF vs. Non-IPF).

Another research [4] added the importance of data preprocessing which involves scaling the dataset for the best model performance and dividing it into training and testing sets. According to this experiment, a DNN model outperformed the best-performing classical machine learning model (J48) with an accuracy rate of 89%. This implies that DNNs may be useful for precise IPF identification. In addition, a web application is being developed as part of the project to deploy the DNN model and maybe make it available for clinical usage.

Some studies [5] examine the prediction of IPF progression using an LSTM network. Important facts of this study consist of Data acquisition which involves making use of a patient's health records, lung function tests, and HRCT images that were taken at various intervals. Preprocessing the data to make sure it is compatible with the LSTM model; this may involve actions such as cleaning the data, normalizing it, and extracting features from HRCT scans. The creation and training of an LSTM model to evaluate sequential data and forecast future illness development based on historical data is known as LSTM Model Development. The paper results that the model has MSE as 33353.5527, RMSE as 183.1217 and MAE as 134.1621.

Overall, the work done by other researchers indicates that both machine learning and deep learning models can be implemented for the prediction with few customizations.

III. Methodology

A. Data-set description

For model training and assessment, OSIC dataset of 200 patients was used. In the training set, baseline Computed Tomography (CT) scan and the entire history of FVC measurements are provided and, in the testing, set baseline CT scan and only the initial FVC measurements are provided. The data are in two forms: textual data as shown in Table 1 and CT scans data as shown in Fig. 1.

Table 1: Sample of Kaggle Dataset [6]

Patient	Week	FVC	Percent	Age	Sex	Smoke-status
ID00007637202177411956430	-4	2315	58.25364	79	M	Ex-smoker
ID00007637202177411956430	5	2214	55.71212	79	M	Ex-smoker
ID00007637202177411956430	7	2061	51.86210	79	M	Ex-smoker
ID00007637202177411956430	9	2144	53.950679	79	M	Ex-smoker
ID00007637202177411956430	11	2069	52.063412	79	M	Ex-smoker

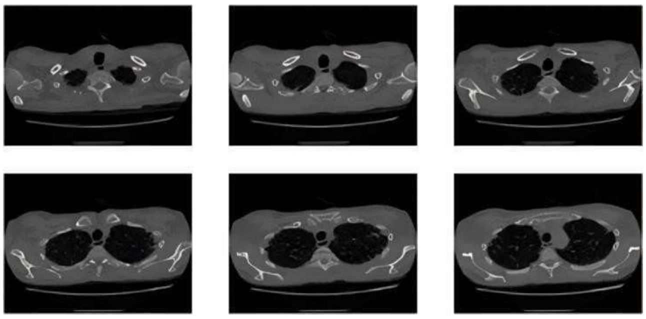

Fig. 1: Sample of Kaggle CT Scans dataset [6]

B. Proposed Models

A multimodal approach is used for making predictions. Fig. 2 explains the approach for predicting FVC using machine learning and deep learning techniques. The Hubber Regression and SVR model use only tabular data to extract relevant features and make predictions while the LSTM model uses both tabular data and CT scans data of the patients to make predictions.

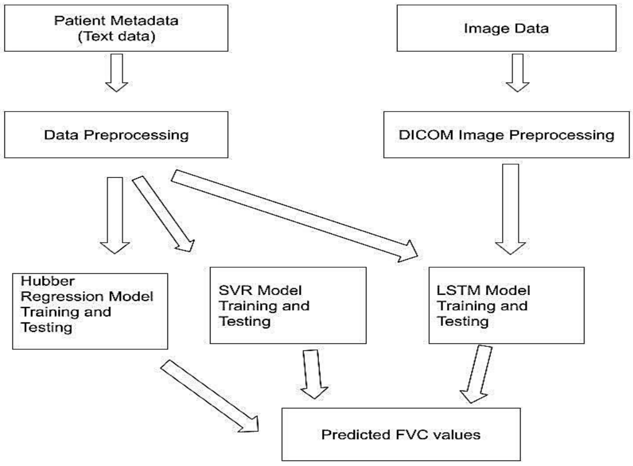
Patient Metadata (Text data)
Image Data
Data Preprocessing
DICOM Image Preprocessing
Hubber Regression Model Training and Testing
SVR Model Training and Testing
LSTM Model Training and Testing
Predicted FVC values

Fig. 2: Schematic diagram of the methodology

The first implemented model is a Hubber Regression model which is a type of robust regression technique. Hubber regression emphasizes the bulk of the data points that show the typical correlations between clinical characteristics and FVC while minimizing the impact of outliers. This makes it possible for the model to more successfully identify the underlying patterns in the "good" data, which leads to more precise forecasts for the majority of patients.

The implementation of model requires data preprocessing which is responsible for cleaning and formatting the data so that potential missing values can be addressed for better modelling. In this, missing values from the table has been removed. Then feature selection was done for selecting potentially relevant clinical features from the provided dataset that have a high impact on prediction. Followed by hubber regression model where actual prediction occurs and the evaluation of the model where performance of the model can be seen. Fig. 3 depicts the difference between the true FVC and the predicted FVC values in the training phase while

Fig. 4 depicts the same difference during the testing phase of t

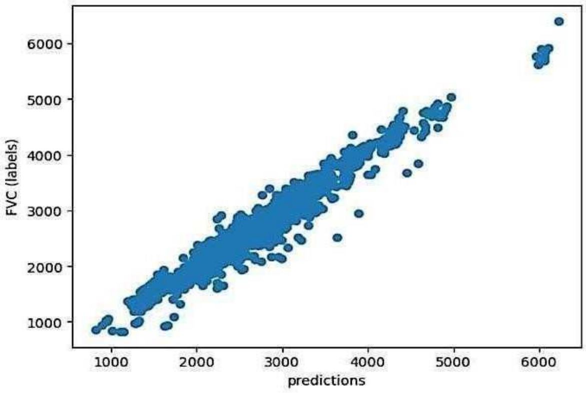

Fig. 3: Plot between true FVC and predicted FVC (training phase)

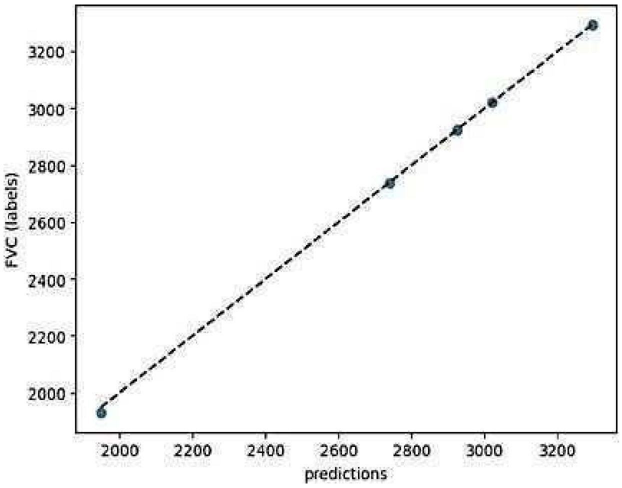

Fig. 4: Plot between true FVC and predicted FVC (testing phase)

The second model is a Support Vector Regression model. SVR is a supervised machine learning algorithm which is known for its robustness to outliers and noise in data. The decision to select this model after working on the Hubber regression model is because of its ability to handle outliers. SVR introduces a margin of intolerance which ignore outliers beyond a certain threshold.

The aim to experiment with this model was because of its ability to capture non-linear relationships between the input features and output or target values. With the given dataset, it is not necessary that all the variables, when they are changed or tweaked, result in a constant change with respect to the target variable. The input variable 'Percent' shows a strong linear relationship with the target variable 'FVC' while 'Age' and 'Weeks' show a non-linear pattern. SVR reduces error rate by fitting the error inside a certain threshold. Regression is performed at a higher dimension feature space. A hyperplane- which is a separation line to distinguish between data points at a higher dimension and their actual dimension is set.

A kernel function is set to map our data points to a higher dimension. Common kernel functions used in the SVR model are the polynomial kernel, Radial Basis Function (RBF) kernel and the sigmoid kernel. In this specific model, it is decided to go with the Gaussian Kernel Radial Basis Function (RBF) because of its ability to capture non-linear relationships. (1) provides the formula to calculate RBF whichmeasures the similarity between two data points by computingthe Gaussian function of the Euclidean distance between them.Additionally, it provides a wide range of flexibility which is important when dealing with medical data as the correlations can be varying from different patients and different stages of the disease.

$$k(x, l^{i}) = e^{-\frac{\|x\to^{-7i}\|^{2} \; -x}{2\sigma^{2}}} \qquad (1)$$

Boundary lines are made to create margin between data points, they are at a distance epsilon from each other and are drawn around the hyperplane. Support Vector are the extreme data points near the boundary line which set the hyperplane. The objective of SVR is to fit as many data points as possible without violating the margin.

SVR also contains two regularization parameters-

- *C Parameter*: The C parameter is used to minimize the training error and to control the trade-off between minimizing the training error and minimizing the complexity of the hyperplane.
- *Gamma Parameter*: The γ function defines the reach of the individual training samples. It controls the smoothness of the decision boundary.

The gamma parameter is set to 'scale' and the C regularization parameter is set to 1.0. The C parameter is used to minimize the training error and to minimize the complexity of our model.

Fig. 5 depicts the plot between true FVC and predicted FVC during the testing phase of the SVR model.

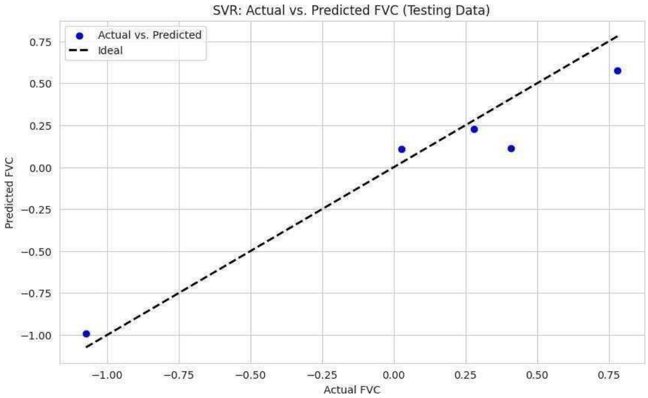

Fig. 5: Plot between true FVC and predicted FVC

The third model is Long Short-Term Memory (LSTM) model which forecast the advancement of pulmonary fibrosis. Due to their unique architecture, LSTM models provide a number of advantages in this arena. Specialized memory cells, or LSTM cells, are what distinguish LSTM models from traditional neural network architectures. These cells have the extraordinary capacity to store and selectively retain information for extended periods of time. The model uses a recurrent neural network (RNN) architecture with LSTM cells specifically integrated to process sequential input efficiently and remember previous observations. The implemented version has a feedback mechanism, which enables iterative prediction refining over several time steps, in contrast to traditional LSTM models. Prediction accuracy is improved through dynamic adaptation to changing patient situations made possible by this special design. The working of model includes: Data Preparation where the dataset is loaded from CSV files containing patient information, including demographics, clinical characteristics, and FVC (Forced Vital Capacity) values. To maintain data integrity, duplicate items in the dataset are eliminated. These entries are found using the patient ID and weeks. To get ready for model training, features including baseline FVC and patient characteristics are computed. Then preprocessing and feature engineering is done where images of lung scans are loaded using the PyDicom library. To get the images ready for input into the LSTM model, image preprocessing methods such lung segmentation, resampling, and Hounsfield Unit (HU) conversion are used. When multimodal data are available, they are incorporated into the dataset to make sure the LSTM design is compatible. The LSTM model architecture is designed using the Keras library. To capture temporal dependencies in the sequential data, various LSTM layer configurations are tested with, including input width, label width, and shift. At the end model training and evaluation is done appropriate optimizers, loss functions, and assessment metrics like MAE and MSE are used for the calculation of the LSTM model.

C. RESULT AND DISCUSSION

In order to predict Forced Vital Capacity (FVC) in patients with suspected pulmonary fibrosis (PF), this study assessed the efficiency of three models: Huber regression, Long Short-Term Memory (LSTM), and Support Vector Regression (SVR). Hubber regression uses a hubber loss which is both MSE and MAE, it is quadratic (MSE) in small error cases and MAE in large error cases. In this case, delta is the hyperparameter that defines the range for the MAE and MSE. Iterative testing helps ensure that the delta value is accurate. Huber regression emphasizes the bulk of the data points that show the typical correlations between clinical characteristics and FVC while minimizing the impact of outliers.

It functions similarly to least squares for lesser errors, but as errors get worse, the penalty progressively rises and extreme outliers receive less weight. This property reduces their impact on the learning process of the model, resulting in more dependable predictions even when outliers are present.

The model has a hubber score of 0.999 which means the model works accurate on nearly 99% of the data. Besides that the model has good Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics as shown in Table 1 and Table 2.

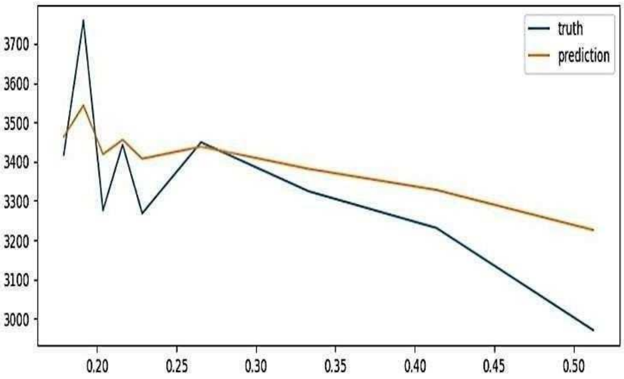

Fig. 6: Plot to show actual values vs predicted values

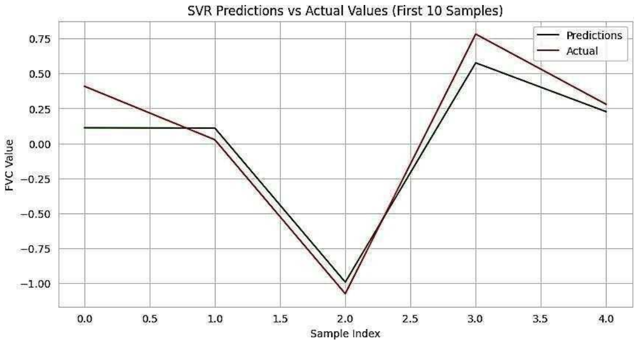

Fig. 7: Line plot to show actual values vs predicted values

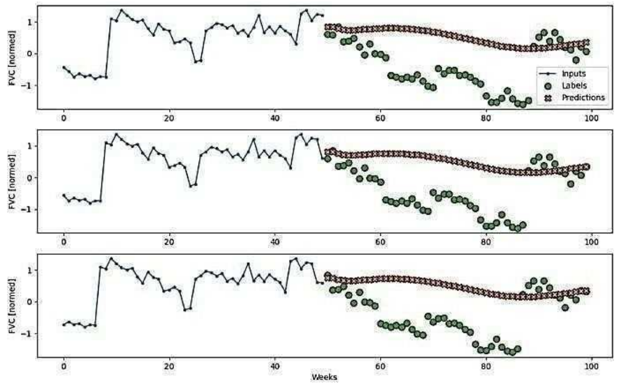

Fig. 8: Multi-Val Performance of AR LSTM

Fig. 6, Fig. 7 and Fig. 8 shows the graphical representation between the actual and predicted values of the three models- the Hubber regression model, the SVR model and the LSTM model respectively. It can be seen from the graphs that all the models closely approximate the actual values.

Principal Results:

As observed in both training and testing phase, SVR has performed well compared to LSTM and Hubber Regression. Overall, all the 3 models have worked well in both the phases.

The comparison of the three models based on the MSE, MAE and RMSE performance metrics is presented in Table 2 for the training values and in Table 3 for the testing values.

Table 2: Comparison in MSE, MAE and RMSE of the three implemented models (training values only)

Model	MSE	MAE	RMSE
Huber Regression	177.350	128.080	2.8200
LSTM	0.845	1.0168	1.2015
Support Vector Regression	0.027	0.1205	0.1650

Table 3: Comparison in MSE, MAE and RMSE of the three models(testing phase only)

Model	MSE	MAE	RMSE
Hubber Regression	7.98	3.57	2.82
LSTM	0.94	1.07	1.39
Support Vector Regression	0.029	0.144	0.171

CONCLUSION

This study evaluated three models for predicting specific values in a given domain.

The results demonstrate the potential of machine learning models in predicting FVC for PF diagnosis and management. The SVR model, leveraging patient metadata, achieved the lowest Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicating superior predictive accuracy compared to the LSTM and Huber regression models. It is also important to note that the LSTM model incorporated both the patient's clinical data (CT scans of lungs) and the patient metadata while the SVR and Hubber models solely relied on the tabular patient metadata to train the model and make predictions.

The Hubber regression model, while robust to outliers, struggled to capture the underlying patterns in the data, resulting in significantly higher prediction errors. It is also noteworthy that although the Huber model's MSE is high, it has a good hubber score of 0.999. These results underscore the importance of selecting appropriate models and incorporating longitudinal clinical information for accurate FVC prediction in PF patients. Different results were obtained from the analysis. The LSTM model, which is renowned for its capacity to identify intricate correlations in sequential data, did not perform better than the other models, despite early predictions to the contrary. When compared to Huber regression, both SVR and LSTM in this particular situation had the lowest MSE, indicating a greater average gap between the predicted and real FVC values.

FUTURE SCOPE

Moving forward, several avenues for future research and application emerge from our study. The study can investigate other interesting strategies, such as ensemble learning, which mixes several models to get predictions that might be even more accurate. Additionally, the integration of additional data may improve model interpretability and performance. It may help to understand the underlying patterns of the disease inn depth.

Collaboration with healthcare professionals to implement machine learning models into routine clinical practice is crucial for evaluating their utility and impact on patient outcomes. Furthermore, longitudinal studies tracking PF progression and treatment response may provide valuable insights into the long-term effectiveness of machine learning- based predictive models.

In conclusion, the study highlights the potential of machine learning models in predicting FVC for PF diagnosis and management. Continued research efforts in model refinement, data integration, and clinical translation are essential for advancing the use of machine learning in PF care, ultimately improving patient outcomes and quality of life.

REFERENCES

[1] A. Yadav, et al., "FVC-NET: An Automated Diagnosis of Pulmonary Fibrosis Progression Prediction Using Honeycombing and Deep Learning," *Wiley*, vol. 2022, no. 12, Article ID 2832400, 2022.

[2] K. Du, et al. "Medium-long term prognosis prediction for idiopathic pulmonary fibrosis patients based on quantitative analysis of fibrotic lung volume." *Respiratory Research,* vol. 23, no. 1, pp. 372,2022.

[3] R. Alawani, "A Machine Learning and Deep Neural Networks Approach to Diagnosing Idiopathic Pulmonary Fibrosis," 2022.

[4] D. Venkatesh, R. Valarmathi, and R. Uma, "An LSTM-based approach for predicting idiopathic pulmonary fibrosis progression," in *AIP Conference Proceedings*, vol. 2464, no. 1, AIP Publishing, May 2022.

[5] S. Y. Ash, et al., "Deep learning assessment of progression of emphysema and fibrotic interstitial lung abnormality," *American Journal of Respiratory and Critical Care Medicine*, vol. 208, no. 6, pp. 666-675, 2023.

[6] OSIC Pulmonary Fibrosis Progression dataset. Kaggle. Retrieved from https://www.kaggle.com/competitions/osic-pulmonary-fibrosis-progression/data.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[8] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 4, pp. 199-222, 2004.

[9] S. Raeymaekers, P. C. Carrillo, A. U. Wells, T. Hoeppner, D. M. Patel, C. I. Silva, and M. Prokop, "Forced vital capacity in idiopathic pulmonary fibrosis," *The European Respiratory Journal*, vol. 56, no. 2, May 2020.

[10] A. S. Oh, et al., "Deep learning–based fibrosis extent on computed tomography predicts outcome of fibrosing interstitial lung disease independent of visually assessed computed tomography pattern." *Annals of the American Thoracic Society,* vol. 21, no. 2, pp. 218-227, Sep 2023.

[11] A. A. Trusculescu, D. Manolescu, E. Tudorache, and C. Oancea. "Deep learning in interstitial lung disease—how long until daily practice." *European radiology*, vol. 30, no. 11, pp. 6285-6292, June 2020.

[12] S. L. F. Walsh, et al. "Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort

study." *The Lancet Respiratory Medicine* 6, vol.6,no. 11, pp.837-845,Sep 2018.

[13] X. Huang, W. Si, X. Ye, Y. Zhao, H. Gu, M. Zhang, S. Wu, Y. Shi, X. Gui, Y. Xiao, and M. Cao. "Novel 3D-based deep learning for classification of acute exacerbation of idiopathic pulmonary fibrosis using high- resolution CT." *BMJ Open Respiratory Research*, vol. 11, p. e002226, Feb 2024.