An LSTM-based approach for predicting idiopathic pulmonary fibrosis progression

Conference Paper *in* AIP Conference Proceedings · May 2022

DOI: 10.1063/5.0082651

3 authors, including:

Valar Mathi
Sri Sairam Engineering college
35 PUBLICATIONS **224** CITATIONS

SEE PROFILE

Uma Ranganathan
Sri Sairam Engineering college
41 PUBLICATIONS **120** CITATIONS

SEE PROFILE

An LSTM-based approach for predicting idiopathic pulmonary fibrosis progression

Cite as: AIP Conference Proceedings **2464**, 060009 (2022); https://doi.org/10.1063/5.0082651
Published Online: 26 May 2022

D. Venkatesh, R. Valarmathi and R. Uma

An LSTM-based approach for predicting idiopathic pulmonary fibrosis progression

D. Venkatesh[1,a)], R. Valarmathi[1,b)] R. Uma [1, c)]

[1]Department of Computer Science Engineering, Sri Sairam Engineering College
Chennai, India

a) Corresponding author: venkatesh27042001@gmail.com
b) valarmathi.cse@sairam.edu.in
c) uma.cse@sairam.edu.in

Abstract. Pulmonary fibrosis is a progressive lung disease that occurs when lung tissues get scarred and damaged. Although this condition cannot be completely treated, early identification and prediction of its progression can assist to keep it under control. Since this disease can occur without any cause it is termed “Idiopathic”. This disease can cause shortness of breath, fatigue, a dry cough, etc., and lead to death if left uncared. The objective of this paper is to use the patient's HRCT images from the CT scanner, forced vital capacity (FVC) assessed with a spirometer, and other patient information like sex, smoking status, and so on to predict the severity of idiopathic pulmonary fibrosis progression in the lungs. Nowadays, Machine Learning plays a significant part in the healthcare sector for predicting and diagnosing various diseases, image segmentation, drug discovery, etc. The LSTM (Long Short Term Memory) model is utilized in this work to predict disease progression. The LSTM is a kind of RNN (Recurrent neural network) that is effectively used for predicting time series data and for sequence prediction problems. This model predicts the future values of FVC measurements through which we can know the patient’s severity of the decline.
Keywords: Idiopathic Pulmonary Fibrosis (IPF), LSTM, RNN, Lung disease, Machine Learning.

INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a lung condition that is persistent and progressive that occurs when the lung tissues get scarred and stiffened. The damage caused by this disease cannot be cured totally. As IPF progresses, the lung loses its ability to hold oxygen and in turn causes breathing difficulty. In severe cases, it may cause shortness of breath and even may lead to a patient’s death. The actual origin of this condition is usually unclear, which is why it is referred to as "idiopathic." After diagnosis, a patient with this condition may expect to live for 2 to 5 years on average [10].In addition, IPF seems to have a much higher death rate. Shortness of breath, a dry cough, tiredness, weight loss, and painful muscles and joints are all signs of IPF. Long-term exposure to pollutants such as silica dust, asbestos fibers, hard metal particles, coal dust, and others, as well as certain medical disorders, radiation therapy, and some medicines, are all recognized causes of IPF.

There are only limited treatments for treating IPF. Furthermore, no proof exists that any medicines can assist to cure IPF since scarring is irreversible once it develops. The prediction of IPF progression during diagnosis is considered to be a difficult task.

Machine learning models are widely used in the healthcare sector for diagnosing various diseases. Pattern recognition and machine learning have the potential to improve the perception and diagnosis accuracy of diseases in the biomedical industry. [8].

High resolution computed tomography (HRCT) is critical in determining the progression of IPF. Also forced vital capacity (FVC) measurements act as an essential parameter in determining the patient’s severity of the decline. The FVC measurements indicating the volume of air exhaled are measured using a spirometer. The FVC measurements are necessary in order to assess lung function.

Early structural variations on HRCT images can be used to predict lung function progression using a quantitative score [2].The Quantum Particle Swarm Optimization and Random Forest (QPSO-RF) algorithm was used to predict growth of disease on HRCT images using only metrics from a single scan for subjects with IPF to predict progressive ROIs at 6 months to 1 year follow-ups using only baseline HRCT scans. [1].Algorithms like CNN, three-dimensional multiscale fuzzy entropy (MFE3D) etc., are also used in diagnosing IPF [3-4]. In addition, one study revealed the first clinical outcomes achieved by applying the differentiation of pulmonary fibrosis using Quantitative LUS Spectroscopy [9].

This paper uses the OSIC dataset for predicting the progression of IPF through the LSTM model.

Proceedings of the 1st International Conference on Modern Functional Materials (ICMFM 2021)
AIP Conf. Proc. 2464, 060009-1–060009-6; https://doi.org/10.1063/5.0082651
Published by AIP Publishing. 978-0-7354-4340-2/$30.00

OSIC DATASET

The Open Source Imaging Consortium (OSIC) is a collaborative effort between industry, academia, and philanthropy which aims to bring together clinicians, radiologists, and computational scientists across the world together to enhance imaging-based therapies.

The OSIC dataset consists of two different files, one containing the folders of HRCT images of the patients and another, a csv file, containing the associated clinical data of the patient.

The HRCT training image folder is divided into 176 subfolders, where each folder consists of collected CT scan images of a particular patient. The csv file contains 2270 records with other clinical information like Age, FVC, Smoking Status, Percent, Week and Sex. Because this is actual medical data, the timing of the first measurement in relation to the CT scan, as well as the duration of the projected time points, may vary per patient [6].

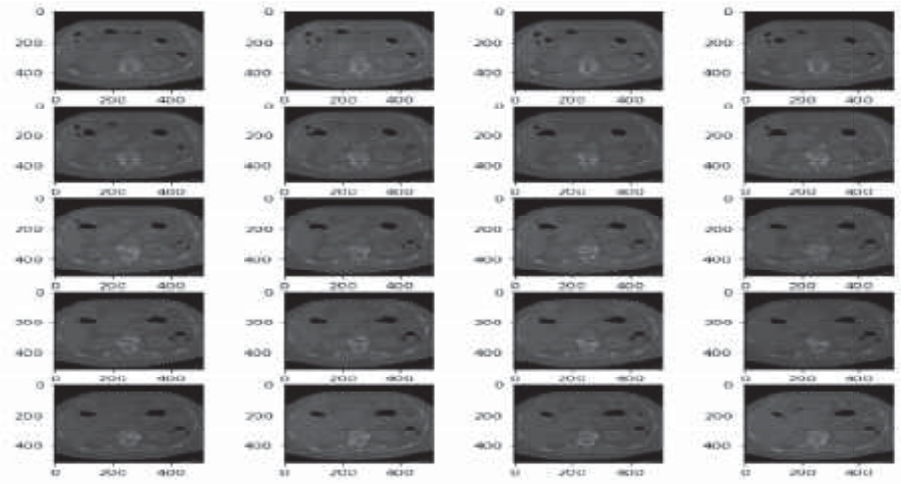


Figure 1.HRCT images of a particular patient

The above Figure 1 shows the sample set of HRCT images of a particular patient.

METHODOLOGY

The suggested solution's process flow is depicted in Figure 2.

FIGURE 2. Process flow of the proposed solution

Data Analysis and Preprocessing

Data Analysis and Preprocessing is an essential step before feeding our data into the model. This step is used to discover useful information, avoid discrepancies in the data and transform it into consistent data.
Correlation Matrix

A Correlation matrix is a symmetric square matrix containing the correlation coefficients between the set of features. The correlation matrix is useful to understand the dependency between a set of features. The value of correlation coefficient lies in the range of -1 to 1.
Figure 3 shows the correlation matrix between Age, Percent, FVC and Weeks. It can be inferred from Fig.3 that FVC and Percent are more positively correlated than other pairs of feature

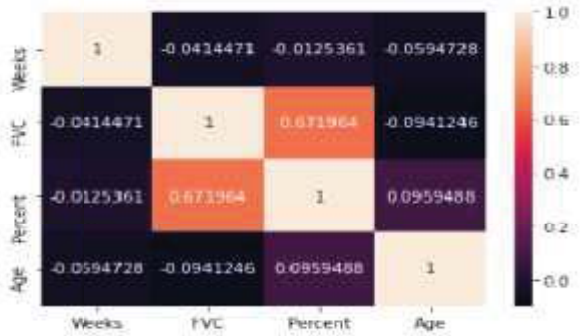
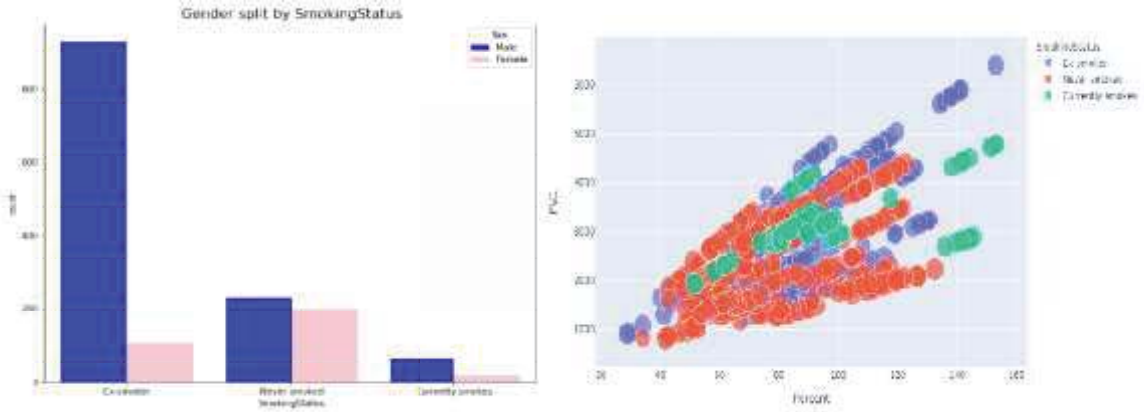


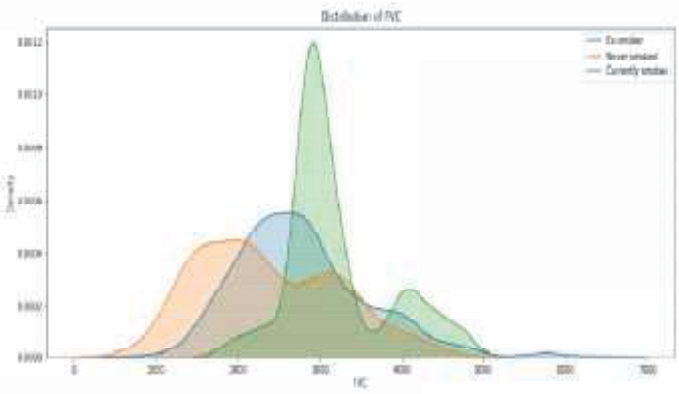
FIGURE 3. Correlation matrix

Visualization Plots

The Visualization plots are used in order to analyse huge amounts of data. Many types of charts like Bar charts, Pie chart, Scatter plots etc., are used for visualization. Visualization plots are useful in better understanding the trends and patterns in the dataset. The Figure 4 depicts some of the visualization plots drawn from the dataset.



(a) (b)



(c)

FIGURE 4.Visualization plots drawn from OSIC Dataset.

Encoding

Encoding is used to convert the categorical features into numeric values. Encoding techniques are used to improve the quality of our model. Here, Label encoding is done on Sex and Smoking Status features of the dataset.

Normalization

Normalization is a technique of rescaling a numeric feature to a common scale without distorting the differences in the range of values.

Here, Z-Score Normalization is applied on Week, FVC, Percent and Age features. In Z-score Normalization, the features are normalized based on their mean and standard deviation.

$$\hat{X}_i = \frac{X_i - \mu}{\sigma} \quad (1)$$

Where $\hat{X}_i$ represents the ith normalized/new value of the feature X, and X_i represents the ith value of featu represents the feature X's mean and σ represents the feature X's standard deviation.

LSTM Model

The LSTM is a form of recurrent neural network that avoids the vanishing gradient problem by training it via Backpropagation across time. LSTM is effectively used for predicting time series data and for sequence prediction problems.

LSTM networks use memory blocks containing gates, instead of neurons, which are connected through layers. A forget gate, an input gate, and an output gate are all included in each memory block. The input gate controls the flow of activations into the memory cell, while the output gate controls the flow of cell activations out into the rest of the network [7].

The Figure 5 depicts the simple architecture of LSTM.

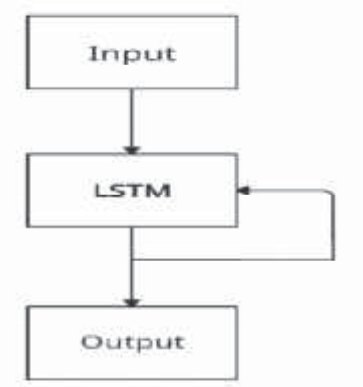


FIGURE 5. Simple LSTM RNN Architecture

The LSTM model was developed and trained with several layers to predict the patient's future FVC measurements.

RESULT

Some of the best metrics that are used to evaluate a regression model are Mean square error (MSE), Root Mean square error (RMSE), Mean absolute error (MAE) and Mean absolute percentage error (MAPE) [5]. All the performance metrics listed above were used to assess the LSTM model's performance. The computed performance metrics of the LSTM model are shown in the table I below.

TABLE I Performance metrics of the LSTM Model

Metrics	Values
Mean squared error (MSE)	33533. 5527
Root Mean squared error (RMSE)	183.1217
Mean absolute error (MAE)	134.1621
Mean absolute percentage error (MAPE)	5.5855

The scatter plot between actual FVC measurements and predicted FVC values is shown in Figure 6.

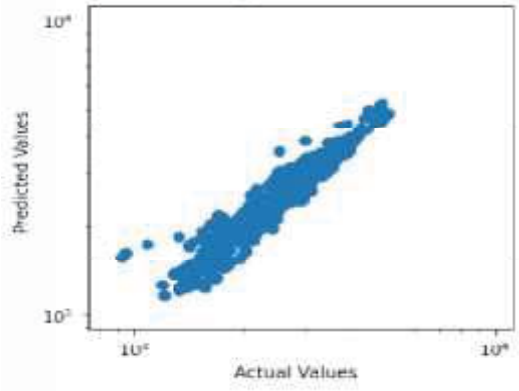


FIGURE 6. Scatter plot between actual FVC measurements and predicted FVC measurements.

CONCLUSION

This paper proposed an LSTM based approach to predict the progression of IPF. The performance of the LSTM model has been analyzed based on various regression performance metrics. It will be effective if this kind of predictive model is deployed in medical sectors for predicting the progression of IPF. It will be useful for the doctors in the early diagnosis of this disease. This may in turn control and prevent the patient's severity of the decline.

REFERENCES

1. Shi, Yu & Wong, Weng&Goldin, Jonathan & Brown, Matthew & Kim, Grace. (2019), "Prediction of Progression in Idiopathic Pulmonary Fibrosis using CT Scans at Baseline: A Quantum Particle Swarm Optimization - Random Forest Approach. Artificial Intelligence in Medicine", 2019 Sep;100:101709. doi 10.1016/j.artmed.2019.101709. Epub 2019 Aug 28. PMID: 31607341.
2. Kim, G.H.J., Weigt, S.S., Belperio, J.A. et al., "Prediction of idiopathic pulmonary fibrosis progression using early quantitative changes on CT imaging for a short term of clinical 18–24-month follow-ups". EurRadiol 30, 726–734 (2020). https://doi.org/10.1007/s00330-019-06402-6
3. Trusculescu, A.A., Manolescu, D., Tudorache, E. et al., "Deep learning in interstitial lung disease—how long until daily practice", EurRadiol 30, 6285–6292 (2020).https://doi.org/10.1007/s00330-020-06986-4.
4. S. F. Gaudêncio et al., "Three-Dimensional Multiscale Fuzzy Entropy: Validation and Application to Idiopathic Pulmonary Fibrosis," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 1, pp. 100-107, Jan. 2021, doi: 10.1109/JBHI.2020.2986210.

5. Botchkarev, A. (2018)," Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology",ArXiv, abs/1809.03006.
6. OSIC organization, “OSIC Pulmonary Fibrosis Progression”, [Online]. Available: https://www.kaggle.com
7. Sak, Haşim / Senior, Andrew / Beaufays, Françoise (2014): "Long short-term memory recurrent neural network architectures for large scale acoustic modeling", In INTERSPEECH-2014, 338-342.
8. Fatima, Meherwar& Pasha, Maruf. (2017), "Survey of Machine Learning Algorithms for Disease Diagnostic.Journal of Intelligent Learning Systems and Applications", 09. 1-16. 10.4236/jilsa.2017.91001.]
9. F. Mento, G. Soldati, R. Prediletto, M. Demi and L. Demi, "Quantitative Lung Ultrasound Spectroscopy Applied to the Diagnosis of Pulmonary Fibrosis: The First Clinical Study," in IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 67, no. 11, pp. 2265-2273, Nov. 2020, doi: 10.1109/TUFFC.2020.3012289.
10. 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.
11. Alex Sherstinsky,"Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network", Physica D: Nonlinear Phenomena, Volume 404,2020,132306, ISSN 0167-2789, https://doi.org/10.1016/j.physd.2019.132306. (https://www.sciencedirect.com/science/article/pii/S0167278919305974).