Predicting idiopathic pulmonary fibrosis (IPF) disease in patients using machine approaches

Sikandar Ali · Ali Hussain · Hee-Cheol Kim*

Inje University, Haeundae Paik Hospital

E-mail : sikandarshigri77@gmail.com

ABSTRACT

Idiopathic pulmonary fibrosis (IPF) is one of the most dreadful lung diseases which effects the performance of the lung unpredictably. There is no any authentic natural history discovered yet pertaining to this disease and it has been very difficult for the physicians to diagnosis this disease. With the advent of Artificial intelligent and its related technologies this task has become a little bit easier. The aim of this paper is to develop and to explore the machine learning models for the prediction and diagnosis of this mysterious disease. For our study, we got IPF dataset from Haeundae Paik hospital consisting of 2425 patients. This dataset consists of 502 features. We applied different data preprocessing techniques for data cleaning while making the data fit for the machine learning implementation. After the preprocessing of the data, 18 features were selected for the experiment. In our experiment, we used different machine learning classifiers i.e., Multilayer perceptron (MLP), Support vector machine (SVM), and Random forest (RF). we compared the performance of each classifier. The experimental results showed that MLP outperformed all other compared models with **91.24%** accuracy.

Key words

Machine learning, idiopathic pulmonary fibrosis, prediction

Ⅰ. Introduction

Idiopathic pulmonary fibrosis is a typical kind of lung disease. This disease causes damage to the lung and subsequently makes it dysfunction. This disease imparts scars on the lungs of the patient, as a result the tissues of lung become stiff and hard [1]. Consequently, a patient can not breath easily and he feels difficulty in breathing. It exacerbates the lung condition continuously and eventually a patient faces serious medical issues. It also distorts the whole architecture of pulmonary which leads to hypoxia, sometime failure in respiration and also cause death [2] The patients who are diagnosed with idiopathic pulmonary fibrosis die within 5 years when they initially diagnosed with this disease [3,4]. The precise diagnosis and the actual reason of this disease has always remained a challenging task. Many researchers tried to find the accurate diagnosis and treatment of IPF during the early stages but there is enough room for improvement [5]. The aim of this paper is to predict the different reasons which cause idiopathic pulmonary disease. There are certain reasons of this disease. The patient may have pneumonia, lung cancer, acute exacerbation of IPF, heart disease, pulmonary embolism and other fatal disease related to lungs. We applied the data driven techniques based on artificial intelligence for the early prediction of this disease based on different disease conditions.

* corresponding author

Ⅱ. Material and Methods

(A). Data Source

We used Haeundae Paik Hospital data with the consent of the patients who visited the hospital. This dataset contains 2425 patients information with 502 different kind features.

(B). Data Preprocessing

The data, we got was in raw form and it contained so many inconsistencies. We applied different data preprocessing techniques to remove the inconsistencies and the redundancies from the data. In this way we made the data fit for machine

learning implementation. There was a total of 502 features in the dataset. We preprocessed the data according to our target values and as the result of data preprocessing, we extracted 18 highly relevant features from the dataset.

(C). Applied machine learning models

We applied different machine learning models, namely Multilayer perceptron, Support vector machine and Random forest. We compared the performance of all the applied machine learning models like for example accuracy, recall, precision etc.

1. Multilayer perceptron (MLP)

Multilayer perceptron is on the most popular machine learning algorithm [6]. It has a vast application in the field of artificial intelligence. It consists of neurons or nodes which are connected with each other and they have weights and also have computation functions which process the data.

2. Support vector machine (SVM)

Support vector machine (SVM) is one kind of supervised machine learning model [7]. It is used for classification, outlier detection and regression. It is very efficient and robust for high dimensional spaces.

3. Random forest (RF)

Random forest [8] is also a popular machine learning algorithm used for regression and classification. It works on the decision tree fashion and predicts by selecting the best possible solution from the tree.

Ⅲ. Results and discussion

We included 2425 records of patients for our study after excluding the data which contained inconsistencies and were not meeting the criterion of our target classification. We chose 18 features for our study after data preprocessing. The dataset was divided into 80% training and 20% testing sets. We applied three machine learning models: Multilayer perceptron, Support vector machine, and random forest. We got some interesting results. We calculated the accuracy, specificity and recall of each model. The results of MLP outperformed both other models. The accuracy of MLP model was 91.24%.

Table 1. Performance measure of applied machine learning models

Classifiers	Accuracy	Precision	Recall
MLP	**0.9124**	0.9166	0.9218
SVM	0.8866	0.8915	0.8962
RF	0.8632	0.8536	0.8723

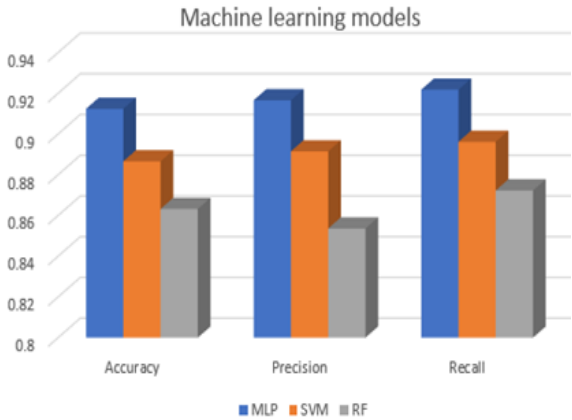

Fig 1. Comparison of accuracy, precision and recall of the machine learning models

In this paper, we used three machine learning classifiers for the prediction of the IPF disease in the patients. A total of 18 highly relevant risk factors were used in the experiment. The accuracy, precision, recall of all the classifiers were shown in Table 1. Figure 1 shows the graphical representation of the applied machine learning algorithms. The performance measures have been compared in the Fig. 1. The experimental results shows that the MLP has highest accuracy as compared to other classifiers.

Ⅳ. CONCLUSION

Idiopathic pulmonary disease is one of the deadliest diseases. Early prediction of this disease is a challenging issue so, in this paper we applied machine learning models to address this global issue. Among the three applied machine learning models, MLP outperformed the other two classifiers and achieved the accuracy of **91.24%**. Consequently, MLP is best in predicting the IPF disease in the patients with high accuracy.

ACKNOWLEDGMENTS

Basic Science Research Program through the National Research Foundation of Korea (NRF), supported by the Ministry of Science, ICT & Future Planning (NRF2017R1D1A3B04032905).

REFERENCES

1. *Idiopathic Pulmonary Fibrosis (IPF)*. Available from: https://.www.webmd.com/lung/what-is-idiopathic-pulmonary-fibrosis
2. Kim, S.Y., et al., *Classification of usual interstitial pneumonia in patients with interstitial lung disease: assessment of a machine learning approach using high-dimensional transcriptional data.* The Lancet Respiratory Medicine, 2015. **3**(6): p. 473-482.
3. Pérez, E.R.F., et al., *Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis: a population-based study.* Chest, 2010. **137**(1): p. 129-137.
4. du Bois, R.M., et al., *Ascertainment of individual risk of mortality for patients with idiopathic pulmonary fibrosis.* American journal of respiratory and critical care medicine, 2011. **184**(4): p. 459-466.
5. Cottin, V. and L. Richeldi, *Neglected evidence in idiopathic pulmonary fibrosis and the importance of early diagnosis and treatment.* European Respiratory Review, 2014. **23**(131): p. 106-110.
6. Gardner, M.W. and S. Dorling, *Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences.* Atmospheric environment, 1998. **32**(14-15): p. 2627-2636.
7. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.
8. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.