In Depth Analysis of Lung Disease Prediction Using Machine Learning Algorithms

Ishan Sen, Md. Ikbal Hossain(✉), Md. Faisal Hossan Shakib(✉), Md. Asaduzzaman Imran(✉), and Faiz Al Faisal(✉)

Department of Electrical and Computer Engineering (ECE), North South University, Dhaka 1229, Bangladesh
{ishan.sen, ikbal.hossain, faisal.shakib, asaduzzaman.imran, faiz.faisal}@northsouth.edu

Abstract. The objective of this examination is to investigate and foresee the Lung Diseases with assistance from Machine Learning Algorithms. The most common lung diseases are Asthma, Allergies, Chronic obstructive pulmonary disease (COPD), bronchitis, emphysema, lung cancer and so on. It is important to foresee the odds of lung sicknesses before it happens and by doing that individuals can be causes and make fundamental strides before it occurs. In this paper, we have worked with a collection of data and classified it with various machine learning algorithms. We have collected 323 instances along with 19 attributes. These data have been collected from patients suffering from numerous lung diseases along with other symptoms. The Lung diseases attribute contains two types of category which are 'Positive' and 'Negative'. 'Positive' means that the person has a lung disease and so forth. The training of the dataset has been done with K-Fold Cross Validation Technique and specifically, five Machine Learning algorithms have been used which are Bagging, Logistic Regression, Random Forest, Logistic model tree and Bayesian Networks. The accuracy for the above mentioned machine learning algorithms are 88.00%, 88.92%, 90.15%, 89.23%, and 83.69% respectively.

Keywords: Machine learning algorithm · Lung diseases bagging · Logistic Regression · Random Forest · Logistic model tree · Bayesian Networks

1 Introduction

Respiratory diseases are among the critical reasons for death around the world. Lung contaminations (for the most part pneumonia and tuberculosis), carcinoma and chronic obstructive pulmonary disease (COPD) together represented 9.5 million passing overall during 2008, one-6th of the overall aggregate, the planet Health Organization evaluates that a proportional four maladies represented one-tenth of disability-adjusted life-years (DALYs) lost worldwide in 2008. Likewise among the 25 most fundamental causes were COPD (positioned sixth in 1990 and ninth in 2010), tuberculosis (positioned eighth in 1990 and thirteenth in 2010) and carcinoma (positioned 24th in 1990 and 22nd in 2010) [1]. Forecast of Lung related malady is a difficult factor looked by specialists and clinics. Right now, the exactness of lung sickness assumes an indispensable job.

A. Bhattacharjee et al. (Eds.): MIND 2020, CCIS 1241, pp. 204–213, 2020.
https://doi.org/10.1007/978-981-15-6318-8_18

Machine Learning methods are widely utilized in medical sectors. Data mining holds great potential to explore the hidden patterns in huge information which will be used for clinical diagnosis. Data mining allow health systems to use data systematically and do the analysis for identifying inefficiencies, best practices that improve care and reduce costs. Detection of lung disease is one among the vital issues and lots of researchers are developing intelligent medical decision support systems to urge better the power of the physicians. Therefore we conducted an enquiry supported machine learning algorithms to seek out the probabilities of occurrence of lung disease before it actually occurs, the only purpose of this analysis is that the identify the accuracy levels of Bagging, Logistic Regression and Random Forest algorithms and to integrate this leads to such a system which is user friendly in order that people can use it whenever they feel convenient.

2 Related Works

In 2002, in the United Kingdom, Martin J Wildman et al. [2] and his team worked on to assess whether the results found in terms of survival were the prognoses among physicians of patients with intense poignancy with obstructive lung malady focused on serious consideration. They have been worked on over 832 patients who were matured 45 years and more with a breathing problem, respiratory flunk, and variation of mental health because of suffering from COPD and asthma, or a conjunction of both. Their yielding is calculated by the physicians and they need around 180 days to predict this outcome. D J Hole et al. [3] and his group worked on determining the connection between the one-second forced expiratory volume (FEV1) and subsequent death. They directed this research on 7058 men and 8353 women who were aged between 45–64 years. They wanted to find fatality rate from all causes, ischemic heart illness, lung cancer, and other cancers, sudden stroke, disease of respiratory organ, and different reasons for death following 15 years of development. They found that 2545 men and 1894 women have died in this period with critical patterns of expanding hazard with lessening FEV1 for both genders, the reasons for death studied afterwards age change, smoking, blood pressure, weight record and social groups are clear. Timor Kadir, Fergus Gleeson [4] had given a review about fundamental lung disease expectation ways to deal with date and feature a portion from their comparative qualities and shortcomings in this paper. AI based lung malignant development expectation models have been offered to help medical practitioners in overseeing unknown or recognized uncertain aspiratory knobs. They have used CNN using trained data, which is a class of deep neural network. While evaluating there were needed to be informed about some limitations whether the patients were smoker or with history of lung nodule. Before testing their system, their works concluded with some related questions for future works that if it were to implement, which output or who should use this system. Anuradha et al. [5] proposed a model, which uses machine learning that is to recognize and analyze the lung disease as ahead of schedule as conceivable which will assist the specialist with saving the patient's life. They used binary classification with different parameter like age, gender, X-ray images and view position. And the expected output is to find whether the patient has lung disease or not. For their particular dataset, they

have used mainly CNN and capsule network algorithm for extraction since their chosen dataset of chest x-ray may have many unnecessary data. This paper portrays how lung maladies were anticipated and controlled, utilizing Machine Learning. With pre-trained model they would improve their accuracy level. Jason et al. [6] and his team in their work they gathered information from aspiratory nodule(s) patients and they were about 1500 and their nodules were till 15 mm identified and distinguished on routinely performed CT chest filters matured 18 years of age or else more established from three scholastic focus in the United Kingdom were utilized for creating risk lamination models. For developing their model, they have used Artificial Intelligence (AI) based dataset. Aiyesha et al. [7] have examined the current clinical system and they have suggested how AI based strategies might be utilized for analysis differentially on Phthisis and Pneumonia. Here they presented a classification model using three machine learning algorithm which are Naïve Bayes, Decision Tree and Random forest. In this paper we also used these algorithms to show different results.

3 Methodology

The examination can be partitioned into four crucial stages which are given as follows:

- Data Collection
- Data Processing
- Data Training
- Application of ML algorithms

3.1 Data Collection

The data for this analysis has been collected from some hospitals in Dhaka. National Institute of Diseases of the Chest and Hospital, National Institute of Cancer Research & Hospital (NICRH) helped us by providing us with the majority of the data. A total of 323 instances and 19 attributes have been collected in which there is information for an individual patient. The dataset contains the lung diseases attribute which is categorized by two types; Positive (have lung diseases) and Negative (Doesn't have any lung diseases) (Table 1).

Table 1. The hospitals which helped us to collect data are mentioned below

1. National Institute of Diseases of the Chest and Hospital
2. National Institute of Cancer Research & Hospital (NICRH)

3.2 Data Preprocessing

In the wake of gathering the information, we did the preprocessing inside the preprocessing stage, we utilized the two unsupervised filters inside the extensively utilized AI stage WEKA 3.8.3 (Waikato Environment for Knowledge Analysis). From the start, we applied the Replace MissingValues filter on our dataset. This replaces all the missing qualities for ostensible and numerical traits utilizing the modes and means. Besides, we've utilized the Randomize filter which replaces the missing data without surrendering a great part of the exhibition.

3.3 Data Training

The preparation of the data has been finished utilizing the K-Fold Cross Validation method of WEKA. It is a resampling procedure to evaluate the forecast model by parting the first dataset into two sections preparing the set and a test set. The parameter K decides the number of groups the dataset will be separated into by rearranging the dataset haphaz--ardly.

3.4 Application of Machine Learning Algorithms

After the training phase, classification has been done using various machine learning algorithms among them Bagging, Logistic

Table 2. Features list

Features	Subcategory	Data distribution
Sex	Male	46.62%
	Female	53.08%
Age	Lowest: 7	Mean ± SD
	Highest: 85	43.412 ± 17.475
Hemoglobin (Hb)	Lowest: 5 gm/dl	11.526 ± 1.491
	Highest: 16.9 gm/dl	
Erythrocyte sedimentation rate (ESR)	Lowest: 2 mm	59.105 ± 33.893
	Highest: 165 mm	
White Blood Cell (WBC)	Distinct: 102	Unique: 22(07%)
Platelet Count (PC)	Distinct: 131	Unique: 43(13%)
Hematocrit (HCT)	Lowest: 9.3%	33.979 ± 4.718
	Highest: 54.6%	
Neutrophils	Lowest: 6%	68.006 ± 14.591
	Highest: 96%	
Lymphocytes	Lowest: 2%	24.035 ± 13.588
	Highest: 90%	
Monocytes	Lowest: 1%	5.168 ± 2.82
	Highest: 17%	
Eosinophil	Lowest: 0%	4.342 ± 2.662
	Highest: 16%	
Basophils	Lowest: 0%	0
	Highest: 0%	
Blood-Glucose-After-Meal	Lowest: 40 mg/dl	106.099 ± 48.443
	Highest: 312.25 mg/dl	
Serum-creatinine	Lowest: 0.3 mg/dl	0.868 ± 0.236
	Highest: 2 mg/dl	
Serum-bilirubin	Lowest: 0.2 mg/dl	1.085 ± 2.249
	Highest: 12 mg/dl	
Serum glutamic pyruvic transaminase (SGPT)	Lowest: 6 u/l	33.049 ± 25.767
	Highest: 178 u/l	
Height	Lowest: 130 cm	160.157 ± 9.608
	Highest: 176 cm	
Weight	Lowest: 20 kg, Highest: 75 kg	50.231 ± 8.989
Cancer-Test-Result {Negative, Positive}	Negative: 307	94.17%
	Positive: 18	5.83%
Lung-Disease-Result	Positive: 285	87.423%
	Negative: 40	12.577%

Regression and Random Forest Logistic model tree and Bayesian Networks outperformed. Therefore, we determined those five algorithms to be our model.

4 Workflow

Figure 1 represents the overall workflow of the entire analysis illustrated in brief. We collected a dataset comprising of 323 instances from hospitals mentioned in Table 2 along with 19 attributes. After that, we preprocessed the data and used the feature selection option in WEKA. The next thing we did was to train the dataset using K-fold Cross-Validation Technique. Afterward, we applied various machine learning algorithms among which three of the above mentioned algorithms stood out. In the end, we concluded our analysis by comparing the performances of the five algorithms.

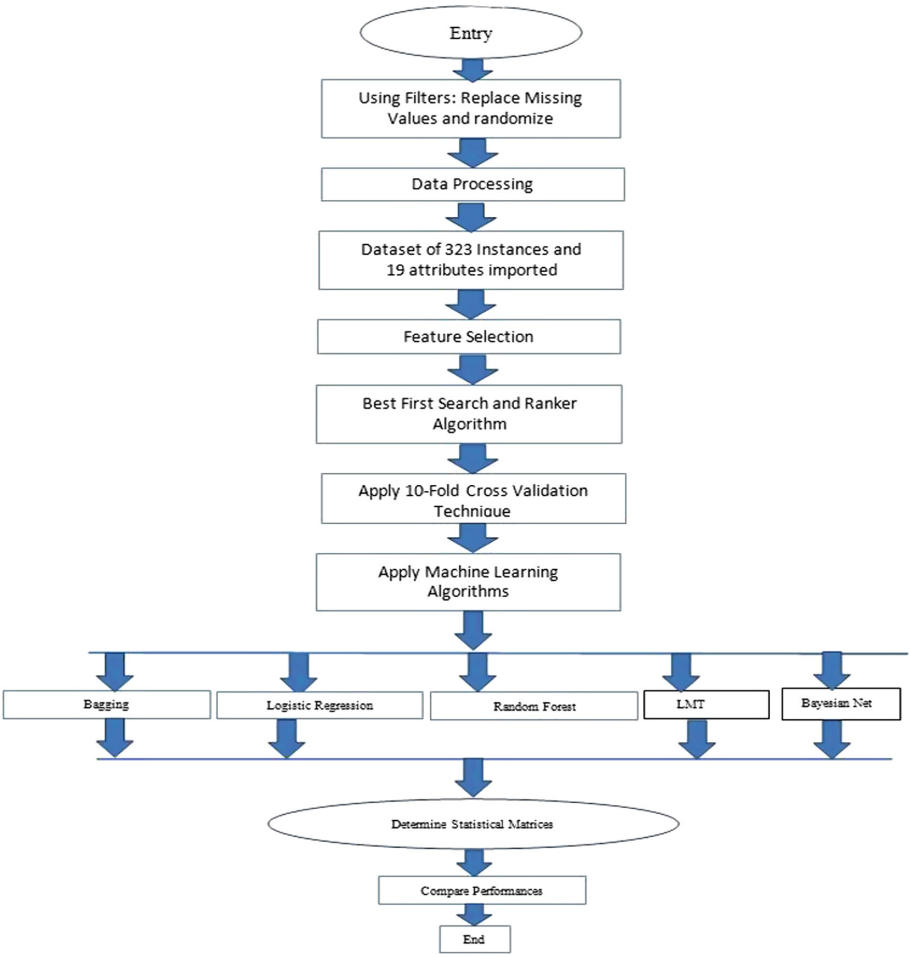

Fig. 1. Flow chart of the overall analysis

Performance Parameters

The outcomes of this analysis are based on the following performance parameters:

A. Seed:

It indicates changing numbers haphazardly and getting an alternate outcome.

B. Correctly Classified Instances (CCI):

The precision of the model relies upon the test information.

$$Accuracy = \frac{\mathrm{Tp} + \mathrm{Tn}}{\mathrm{Tp} + \mathrm{Tn} + \mathrm{Fp} + \mathrm{Fn}}$$

Here, *Tp* = True Positive, *Fp* = False Positive, *Fn* = False Negative, *Tn* = True Negative.

C. Kappa Statistics (KS):

The Kappa Statistics is used to quantify the proclamation among predicted and observed provisions of a dataset.

$$K = \frac{\mathrm{R0} - \mathrm{Re}}{1 - \mathrm{Re}}$$

Here, R_0 = Relative watched understanding among raters, *Re* = Theoretical likelihood of chance assertion.

D. Mean Absolute Error (MAE):

Without assessing the sign it normalizes the size of individual mistakes.

$$MAE = \frac{|\mathrm{p1} - \mathrm{b1}| + \cdots + |\mathrm{pn} - \mathrm{bn}|}{n}$$

Here, p = Predicted Value, b = Actual Value.

E. Relative Absolute Error (RAE):

It is the absolute error with the similar manner of normalization.

$$RAE = \frac{|\mathrm{p1} - \mathrm{b1}| + \cdots + |\mathrm{pn} - \mathrm{bn}|}{|\mathrm{b1} - \mathrm{b}| + \cdots + |\mathrm{bn} - \mathrm{b}|}$$

F. Specificity:

It is the proportion of patients without the disease who test negative.

$$Specificity = \frac{\mathrm{Tn}}{\mathrm{Fp} + \mathrm{Tn}}$$

G. Precision (PRE):

It is denoted as PRE ($PRE = \frac{\mathrm{Tp}}{\mathrm{Tp} + \mathrm{FP}}$).

H. Recall (REC):
It is denoted as REC and the cohesion between PRE and REC is MCC.

$$REC = \frac{\mathrm{Tp}}{\mathrm{Tp} + \mathrm{Fn}}$$

I. F-Measure:
It is denoted by FM.

$$FM = 2 \times \frac{\mathrm{PRE} \times \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}} = \frac{2 \times \mathrm{Tp}}{2 \times \mathrm{Tp} + \mathrm{Fp} + \mathrm{Fn}}$$

5 Results Analysis

The examination was finished in 3 seeds utilizing every calculation for 2 classes titled as Positive and Negative. In Figs. 2 and 3, where the x-axis shows to the False positive rate (*Fp*) and the y-axis connotes the True positive rate (*Tp*). *Tp* is likewise perceived as Recall and it is utilized to quantify the real level of accurately ordered cases. *Fp* is the proportion between the number of negative cases that are inaccurately delegated positive and the quantity of really negative examples. (0, 1) a point on the diagram is where the classifier gives the ideal outcome for example it orders the genuinely positive and negative cases effectively.

In Fig. 3a, 3b given below illustrates the percentage of correctly classified instances and incorrectly classified instances respectively from 5 different algorithms. From Fig. 3a, it has clearly seen that the Random Forest (RF) gives the highest accuracy results whereas Bayes Net gives the lowest accuracy; the figure is approximately 90.1538% and 83.6929% successively. On the other hand, Logistic Regression (LR) gives us nearly 88.9231% which is higher than Bagging which gives us almost 88% accuracy. However, the logical model tree (LMT) is slightly higher than LR and the figure is around 88.2308%. From Fig. 3b it can be said that Bayes Net has given us the highest incorrect instances and the percentage is around 16.30 whereas the RF and LMT give us almost lowest incorrect instances and the figure are almost 10%. Table 3 represents the comparison of five different algorithms on various evaluation matrix. It can be observed that the KS value for Bagging, LR, RF, LMT, Bayes Net are 0.2821, 0.3803, 0.501, 0.4052, 0.2974 respectively.

The MAE for Bagging, LR, RF, LMT, Bayes Net are 0.1785, 0.1592, 0.1551, 0.197 successively whereas the Root Mean Squared Error are 0.3102, 0.3038, 0.2888, 0.3015 and 0.3258 respectively for these algorithms. In terms of RAE, it can be said that the Bayes Net gives the highest percentage of error, which is 90.4575% whereas

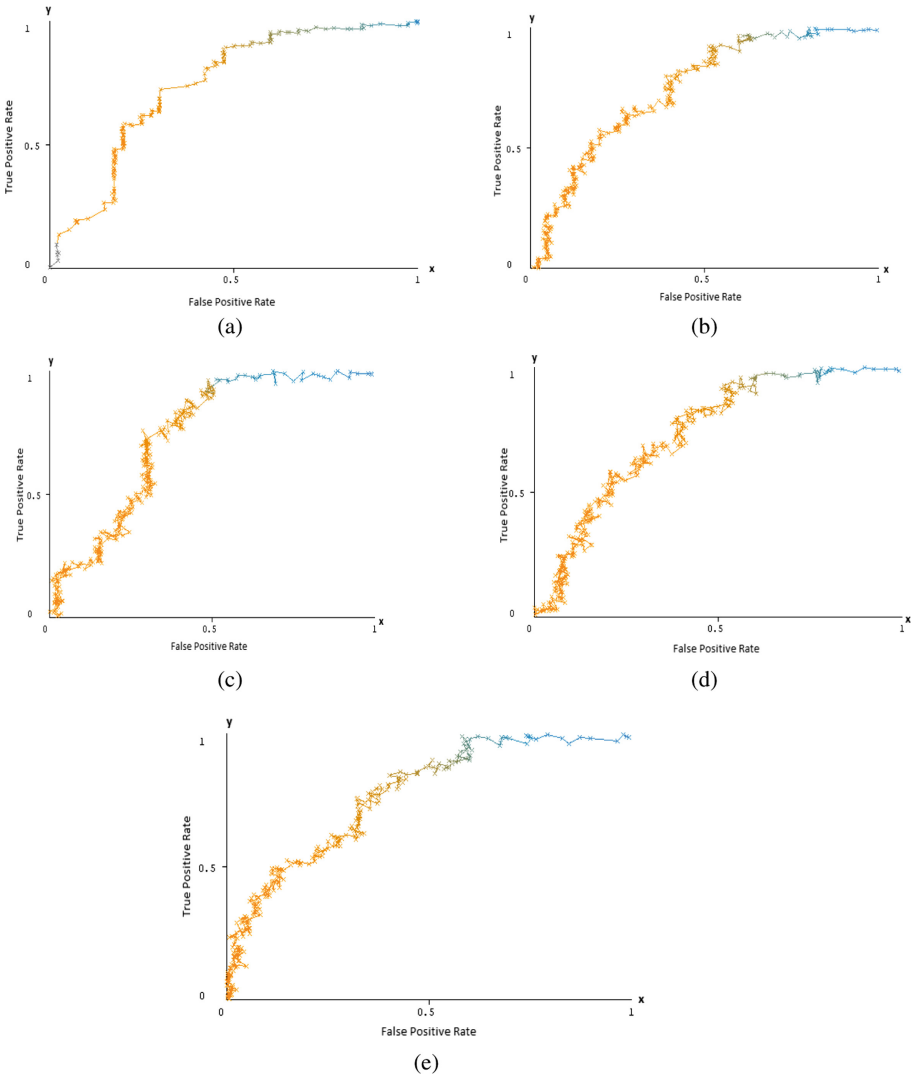

Fig. 2. (a) BAG curve for positive class (b) LR curve for positive class (c) RF curves for positive class (d) LMT curve for positive class (e) Bayes net curve for positive class

the lowest proportion of these errors is 70.1852% percentage, which is for RF. On the other hand, this proportion is slightly lower which is for LMT and the figure is 71.1941%. If it comes to Bagging the error is 81.9378% which is the second-lowest error after Bayes Net. Now if we talk about the Root Relative Squared Error then the Bayes Net gives 99.1762% whereas RF gives 87.9163%. On the other hand, Bagging, LR, KMT give 94.4328%, 92.4839%. 91.7746% respectively.

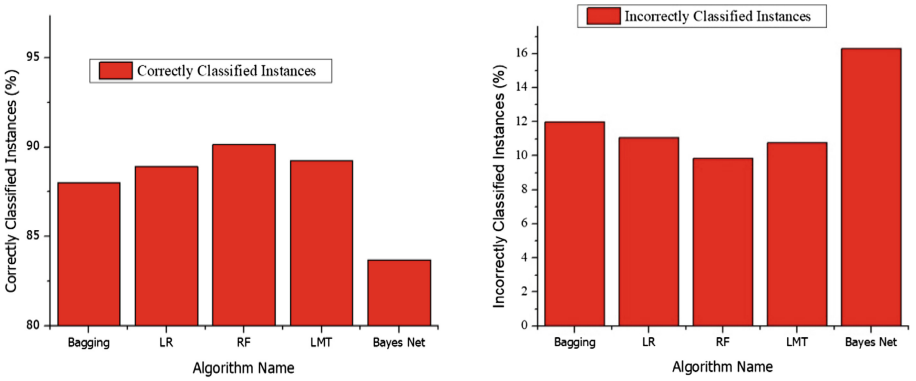

Fig. 3. (a) Correctly classified instances (b) Incorrectly classified instances

Table 3. Statistical metrics comparison for seed 1

Evaluation Metrix	Bagging	Logistic Regr. (LR)	Random Forest (RF)	LMT	Bayes Net
Kappa Statistic	0.2821	0.3803	0.501	0.4052	0.2974
Mean absolute Error	0.1785	0.1592	0.1529	0.1551	0.197
Root Mean Squared Error	0.3102	0.3038	0.2888	0.3015	0.3258
Relative absolute Error	81.93%	73.09%	70.18%	71.19%	90.45%
Root Relative Squared error	94.43%	92.48%	87.91%	91.77%	99.17%
TR Rate (Weighted Avg.)	0.880	0.889	0.902	0.892	0.837
FR Rate (Wei. Avg.)	0.662	0.574	0.444	0.552	0.517
Precision (Wei. Avg.)	0.856	0.873	0.894	0.878	0.849
Recall (Wei. Avg.)	0.880	0.889	0.902	0.892	0.837
F-Measure (Wei. Avg.)	0.861	0.877	0.897	0.881	0.842

6 Conclusion

In conclusion, our study shows that lung diseases can be correctly classified using the machine learning techniques. However, obtaining the real time data was one of the primary concerns that we had faced at the initial stages. In addition to that, we could not able to get similar data from existing works to compare our results. However it is worth to mention that our dataset has many attributes, which is rare to find from online sources. Among the five algorithms, Random Forest gives the best performance than LR, Bagging, LMT and Bayes Net. The accuracy level are 88.00%, 88.9231%, 90.1538%, 89.2308% and 83.6929% respectively. In addition to work, future researches like, study with deep learning methods like - Neuron Advanced Ensemble Learning, Fuzzy Inference System, and Convolution Neural Network would be useful and beneficial.

References

1. The burden of lung disease. https://www.erswhitebook.org/chapters/the-burden-of-lung-disease/
2. Wildman, M.J., et al.: Implications of prognostic pessimism in patients with chronic obstructive pulmonary disease (COPD) or asthma admitted to intensive care in the UK within the COPD and asthma outcome study (CAOS): multicenter observational cohort study (2007). https://doi.org/10.1136/bmj.39371.524271.55
3. Hole, D.J., et al.: Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study (1996). https://doi.org/10.1136/bmj.313.7059.711
4. Kadir, T., Gleeson, F.: Lung cancer prediction using machine learning and advanced imaging techniques (2018). https://doi.org/10.21037/tlcr.2018.05.15
5. Gunasinghe, A.D., Aponso, A.C., Thirimanna, H.: Early prediction of lung diseases. Conference Paper, March 2019
6. Oke, J.L., et al.: Development and validation of clinical prediction models to risk stratify patients presenting with small pulmonary nodules: a research protocol. Diagn. Progn. Res. **2**, 22 (2018)
7. Sadiya, A., et al.: Differential diagnosis of tuberculosis and pneumonia using machine learning. Int. J. Innov. Technol. Explor. Eng. (IJITEE) **8**(6S4), 245–250 (2019). ISSN 2278-3075