ISSN 2047-3338

Spirometric Data Classification on the Basis of Statistical Data Mining

Ms. Prachi D. Junwale, Prof. A. W. Bhade and Dr. P. N. Chatur

***Abstract*—Respiratory disease is a medical term that encompasses pathological conditions affecting the organs and tissues. A lungs disease affects health condition of many people. Lungs diseases can be curable in early detection. Spirometry is valuable for diagnosing specific lung disorders as well as detecting lung disease at an early stage. Spirometry (the measuring of breath) is the most common of the pulmonary function tests and uses an instrument called a spirometer to measure the amount of air entering and leaving the lungs. This test is often used to help doctors diagnose and determine the severity of various respiratory diseases. The output of spirometry is in the form of graphs i.e. flow-volume loop and volume-time curve. This graph gives various parameters that are used for spirometry modelling. In this paper, various pulmonary diseases such as obstructive, restrictive and mixed lung disorders are classified using statistical data mining approach. This classification helps a physician in diagnosis process of various diseases. This approach is used to increase the efficiency of classification.**

***Index Terms*— Lung Disorders, Modelling, Spirometry and Statistical Data Mining**

I. INTRODUCTION

LUNG disease is any disease or disorder that occurs in the lungs or that causes the lungs to not work properly [1]. Early and correct detection of respiratory problems is essential in many related treatments. Measurement of respiratory function is necessary to detect different pulmonary abnormalities. Various techniques are used to detect pulmonary problem such as pulmonary function tests (PFT). PFT measure how efficiently lungs perform. PFTs help physicians to evaluate lung function. Spirometry is the most widely used pulmonary function test.

A. Spirometry

Spirometry is a simple breathing test, which can be done in physician's office that measures the amount of air you can blow out after you have taken in the deepest breath you can [5]. Spirometry is valuable for diagnosing specific lung disorders as well as detecting lung disease at an early stage. In summary, spirometry can:

- Identify the presence of lung disease.
- Help the physician assess the severity of the patient's lung disease.
- Identify pulmonary disease symptoms and the degree of disability.
- Assist in the management of patients with lung disease.
- Provide early detection of pulmonary disease.
- Assist in convincing patients to quit smoking.
- Help the physician assess the effects of therapy or medications.

The result of spirometry test gives various parameters such as FVC (Forced Vital Capacity) and FEV1 (Forced Expiratory Volume), which is mainly used for classification of various respiratory diseases [5]. FVC is the total volume of air you can forcefully blow out. It is an assessment of the size of your lungs, how well your lungs expand and contract, and how well the air passages open and close. FEV1 is the volume of air that you can blow out in the first second of exhalation and FEV1/FVC is the ratio compares the volume of air expelled in the first second to the total volume expelled.

Two curves are shown after the spirometry tests that are flow-volume loop and volume-time curve [5]. The various parameters such as PEF (peak expiratory flow), FIT (forced inspiratory time.) and FET (forced expiratory time) can be determined by using flow-volume loop and volume-time. These parameters are used to determine the values of model parameters. PEF is the maximal expiratory flow rate achieved and this occurs very early in the forced expiratory manoeuvre. FET is the forced expiratory time. FIT is the forced inspiratory time.

There are various fields that are applied on spirometric data for performing post-processing research in respiratory disease analysis area. Such field includes pattern recognition approach, image processing, genetic algorithm, and artificial neural network [8]. In this paper, we will concentrate on data mining approach. For classification of spirometry data, initially spirometry modeling is done on the basis of spirometric parameters and then using data mining method [4], classification of various diseases is performed.

Ms. Prachi D. Junwale is a student of M. Tech (IV Sem) with the department of Computer Science and Engineering , GCOEA, Amravati (M.S.)-444604, India (Tel. No. 9960428440; email: prachi.junwale@gmail.com)

Prof. A. W. Bhade is a Head of Department of Information Technology, GCOEA, Amravati (M.S.) – 444604, India

Dr. P. N. Chatur is a Head of Department of Computer Science & Engg. GCOEA, Amravati (M.S.)-444604, India

B. Spirometry Model

Spirometry lung functioning test is based on the flow measurement. Maximal expiration and the maximal inspiration are the essential things during spirometry test. The breath–in flow and breath–out flow are different from each other. The flow–time curves can be modeled on the basis of the measurements [6]. Fig. 1 shows spirometry model that is used to describe the airflow during the maximal inspiration *Qin* (*t*) and during the maximal forced expiration *Qout*(*t*) [7]. The spline functions were used for the modeling process [6]. The spline functions consist of piecewise lengths of regression functions that give the best fit to localized sections of the data. Statistica data miner concept is used for regression function. The flow-time regression functions were fitted to the flow–time curve based on the spirometry measurements. The maximal inflow *Qmaxin(t)* has been modeled with the regression function given as in (1).

$$Q_{maxin}(t) = A_{in} \cdot \sin(\omega \cdot t), \quad t0 \le t \le t1. \qquad (1)$$

where A*in* is amplitude, ω is pulsation, (t0, t1) the time of breath–in [1]. The maximal outflow *Qmaxout(t) has* been modeled by two exponential spline functions given in (2).

$$Qmaxout(t) = \begin{cases} A^1_{out}(1 - e^{-B1out\, t}), & t1 < t \le t2. \\ A^1_{out} \cdot e^{-B2out\, t}, & t2 < t \le t3. \end{cases} \qquad (2)$$

where *A*1*out* , *A2out* , *B*1*out* , *B2out* are the parameters of regression function [1]. Parameters of the model, which is described by above two equations, are arranged in the vector p[p1,p2,…,pi]= [*Ain,* ω, *A1out, B1out, A2out, B2out*], where *i* = 1,2...6 and the spirometry measurement results are arranged in the vector of measurement y = [*y*1 , *y*2 ,..., *yj*]= [*FVC*, *PEF*, *PIF*,*TPEF*, *FET*], where *j* = 1,2...5. The parameters of model were calculated by using the airflow measurements Q(t). The flow–time regression functions, given by the above two equations, were fitted to the flow–time curve Q(t) based on spirometry measurements.

II. Methods

In this paper, statistical data miner is used to classify restrictive, obstructive and mixed pulmonary diseases. Obstructive lung disease is a category of respiratory disease characterized by airway obstruction. It is generally characterized by inflamed and easily collapsible airways, obstruction to airflow, problems exhaling and frequent office visits and hospitalizations [8]. Restrictive lung diseases are characterized by reduced lung volume, either because of an alteration in lung parenchyma or because of a disease of the pleura, chest wall, or neuromuscular apparatus. In mixed pattern, respiratory system has both obstructive and restrictive abnormalities [8].

The traditional method of classification of a pulmonary disease is done with the help of three respiratory parameters (ordered in measurements vector) y = [*y*1, *y*2, *y*3] = [*FEV*1, *FVC*, *FEV*1%*FVC*]. The reference value of the particular parameter *yj*, j = 1, 2, 3 given in (3) for a patient of age

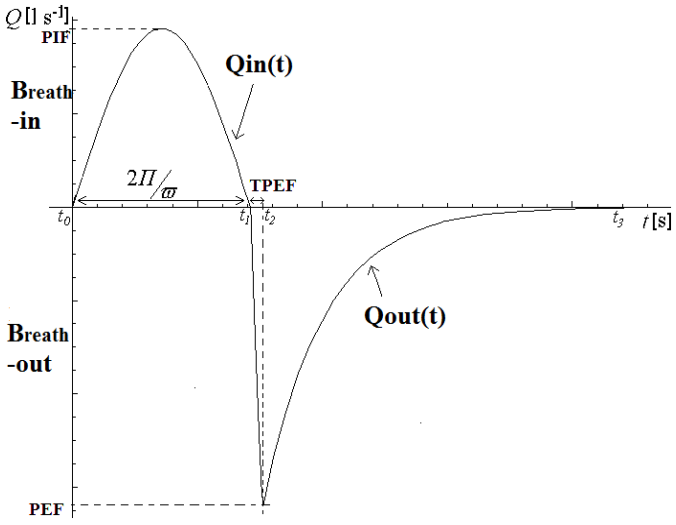

Fig. 1. The flow–time curves: *Qin* (*t*) and *Qout* (*t*)

A[years] and height *H*[*meters*] is obtained using Table I.

$$yj = R^j_1 \cdot H + R^j_2 \cdot A + R^j_3, \qquad j = 1,2,3 \qquad (3)$$

where y1=FVC, y2=FEV1, y3=PEF or *FEV*1%*FVC*.

The norms provide the values of the coefficients R_1^j, R_2^j, R_3^j, for each parameter. The coefficients for *FVC, FEV*1 and *PEF*, published by the European Respiratory Society [2]–[3] are presented in Table I. According to the norms, the age range is from 25 to 70 years and the height range is from 1.55 to 1.95 meters (for men) and from 1.45 to 1.80 meters (for women). The respiratory parameters demonstrate stronger dependence on the height (large absolute value of R_1^j) than on the age (small absolute value of R_2^j).

The individual spirometry results vary depending on condition of the ventilation mechanism. In general, the results close to 100% of the nominal values are interpreted as normal. The results that differ 20% and more from the nominal value are considered abnormal.

TABLE I

THE COEFFICIENTS R_1^j, R_2^j, R_3^j OF THE REFERENCE EQUATION (3) FOR THE RESPIRATORY PARAMETERS FVC, FEV1 AND PEF ACCORDING TO THE EUROPEAN RESPIRATORY SOCIETY

Coefficients	R_1^j	R_2^j	R_3^j
For men			
FVC, *j* = 1	5.760	– 0.026	– 4.340
*FEV*1 *j* = 2	4.300	– 0.029	– 2.490
PEF, *j* = 3	6.140	– 0.043	0.150
For women			
FVC, *j* = 1	4.430	– 0.026	– 3.280
*FEV*1, *j* = 2	3.950	– 0.025	– 2.600
PEF, *j* = 3	5.500	– 0.030	– 1.110

TABLE II

THE OBSTRUCTIVE, RESTRICTIVE AND MIXED DISEASE ARE DIAGNOSED BY DECREASE OF SPIROMETRY PARAMETERS BELOW THEIR NOMINAL VALUES

	Obstructive	Restrictive	Mixed
*y*1	≤ 80% *y*1nom	≤ 80% *y*1nom	≤ 80% *y*1nom
*y*2	*y*2nom	≤ 80% *y*2nom	≤ 80% *y*2nom
*y*3	≤ 70% *y*3nom	*y*3nom	≤ 70% *y*3nom

The decrease of the measurements below their nominal values y*jnom*, *j* = 1, 2, 3 is diagnosed as an obstructive, destructive and mixed disease. Table II shows the details.

The question arises whether the model parameters are accurate enough to be useful for diagnosis purposes. This problem is solved by using spirometric modelling technique. This technique uses the model parameters for classification of respiratory diseases. This is done by using statistica data mining. The result of this classification is referred by physician for further analysis of diseases.

Data mining is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data [4]. The process typically involves various stages, starting with the preparation and "cleaning" of the data, and culminating the identification of models, decision trees, or "important variables" that explain the business or research problems under investigation. *STATISTICA Data Miner* provide a rich tool-box of tools for successfully extracting useful results from structured or unstructured data , and for presenting those results in a meaningful and easy-to-understand manner so that quick decisions and actions can be put in place to turn the extracted "nuggets-of-information" into "nuggets-of-gold" for your business or research project [4]. Descriptive statistics is one of the important features of statistical data mining [4]. It describes how simple summary statistics will help us to further understand our data set. Backed with simple graphical tools, these summaries form the basis of practically every quantitative analysis of data. One of the important method of descriptive statistics is "Breakdown and One-Way ANOVA" that involves dividing ("splitting") the data set into categories in order to compare the patterns of data between the resulting subsets [3]. This technique is used for comparison between measured parameter's value and its nominal parameter's value which is basically used for classification. The nominal model parameters p*nom* were estimated on the basis of y*nom*. The changes of respiratory parameters simulate various health conditions: an obstructive, restrictive and mixed disease. Then, on the basis of simulated diseased cases, the model parameter estimates were calculated: p*obstr*, p*rest* and p*mix* [6]. For each parameter *pi*, i =1,2...6 a relative change in the parameter value was calculated as given in (4).

TABLE III

THE SUSCEPTIBILITY OF THE MODEL PARAMETERS TO VARIED HEALTH CONDITIONS

P*i*	$\frac{Pobstr,i-pnom,i}{pnom,i}$ [%](std. dev)	$\frac{Prest,i-pnom,i}{pnom,i}$ [%](std. dev)	$\frac{Pmix,i-pnom,i}{pnom,i}$ [%](std. dev)
A-in	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
$\tilde{\omega}$	0.000 (0.000)	25.000 (0.001)	25.000 (0.001)
Aout 1	0.000 (0.000)	78.310 (0.172)	78.310 (0.172)
Bout 1	0.000 (0.000)	-25.591 (0.056)	-25.591 (0.056)
Aout 2	12.813 (0.007)	20.710 (0.038)	29.059 (0.034)
Bout 2	12.664 (0.018)	33.966 (0.011)	41.928 (0.024)

p*i* = p*i*(FEV1,FVC,FEV1%FVC) = p*i*(y1,y2,y3)

$$\Delta p_i / p_i = \sum_{j=1}^{j} \delta p_i(y1, y2, y3)/\delta yl \times \Delta yj/yj \qquad (4)$$

The Table III presents the mean values of Δ*pi/pi* and their standard deviations, *i* = 1 to 6 obtained for the whole range of the age and height and for both genders [6].

Table III shows that the obstructive disease causes a noticeable change only in two model parameters i.e., *Aout*2 and *Bout*2. The restrictive lung change causes changes in the model parameters $\tilde{\omega}$, *Aout*1, *Bout*1, *Aout*2, and *Bout*2 respectively. The changes of the model parameters $\tilde{\omega}$, Aout1 and *Bout*1 are the same for both restrictive and mixed diseases. The susceptibility of *Aout*2 and *Bout*2 to mixed changes is more significant than the susceptibility to restrictive changes.

III. Conclusion

The various spirometric parameters such as FVC, FEV1, PEF, FIT and FET are used to determined spirometric model parameters. The spirometric model is designed on the basis of model parameters using spline functions. The obstructive, restrictive and mixed lung disorders are classified using statistical data mining approach. The model parameter estimates lung's health condition. The various model parameters such as $\tilde{\omega}$, *Aout*1 and *Aout*2, when compared to the traditional diagnostic parameters, give new information concerning breathing conditions. Statistical analysis is used to determine the influence of health conditions on the model parameter values during classification. This is done using ANOVA technique. This approach increases the efficiency of classification.

REFERENCES

[1] P. Quanjer, J. Tammeling, J. Cotes, "Long volumes and forced ventilatory flows: report of working party, standardization of lung function tests", *Eur. Respir. Journal,* vol. 6, 1993, pp. 5-40.

[2] P. Quanjer, M. Lebowitz, I. Gregg, "Peak expiratory flow: conclusions and recommendations of a Working Party of the European Respiratory Society. Official ERS Statement", *Eur. Respir. J.*, vol. 10, 1997, pp. 2-8.

[3] J. Kowalski, A. Koziorowski, L. Radwan, Ocena czynności płuc w, chorobach układu oddechowego, "The estimation of the lung function at the pulmonary diseases", Warszawa, Borgis, 2004.

[4] e-handbook of Statistica http://www.statsoft.com, 2005.

[5] e-handbook of breeze-suite http://www.medgraphics.com, 2005.

[6] Renata Kalicka, Wojciech Słomiński and Krzysztof Kuziemski, "Modelling of Spirometry: diagnostic usefulness of model parameters", *IEEE International conference on Computer as a Tool, 2007.*

[7] R. Kalicka, W. Słomiński, K. Kuziemski "Modelling of the spirometry measurements", *XV Krajowa Konferencja Naukowa Biocybernetyka i Inżynieria Biomedyczna,* Wrocław, 2007, pp. 100-104.

[8] S. Jafari, H. Arabalibeik and K. Agin, "Classification of normal and abnormal respiration patterns using flow volume curve and neural network", *IEEE international conference on Information Technology*, 2009.