Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 167 (2020) 313–322

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2019)

Random Subset Feature Selection and Classification of Lung Sound

Don.S[a,*]

[a]*School of Computer Science and Engineering, TIFAC CORE in Automotive Infotronics, VIT, Vellore, Tamil Nadu, India*

Abstract

The lung sounds produced by a human convey valuable information about the health of the respiratory system, and these signals are complex in nature. In this paper, a study was conducted to find the importance of feature selection from these signals for the purpose of classification. Feature selection is performed using two different approaches: RSFS and SFS. The experiment was conducted on a dataset of 85 samples using the (SVM, KNN, and Naïve Bayes) classifiers. The computational results obtained are promising, and the proposed feature selection techniques show better performances in terms of Precision, Recall, and F-Measures.

Keywords: Feature Selection, Fractal Dimension, RSFS, SFS, Random Sampling, Classification;

1. Introduction

Recent studies show that 20.4 million adults above the age of eighteen in the U.S have asthma. Lung sounds constitute a relevant source of information for the study of pulmonary pathology. Doctors use various techniques to study lung sound characteristics. The stethoscope is the instrument most commonly used to determine pathological conditions from human lungs. The accuracy of characterizing sound depends on the practical experience of the physician. Other methods such as X-rays are cost effective, but the radiation is harmful for the body. Thus, it is advisable to have an automated system that can efficiently analyze pathological conditions using lung sounds. Several aspects of respiratory sounds have been observed using machine learning techniques. These observations progress through multiple stages: the pre-processing stage is first, followed by feature extraction and classification. In the pre-processing stage, a method such as resampling is performed to reduce the heart sound effect. During feature extraction, one can extract important information that represents the given data, so that in the classification stage we can label this data as belonging to different class groups. Lung sounds are non-linear, non-stationary [1], and complex in nature. Auscultation of lung sounds by physicians helps in the diagnosis of pulmonary diseases. Lung sounds can be broadly grouped into two categories: normal sounds, and abnormal sounds. In normal breathing, the intensity of sound during

* Corresponding author. Tel.: +91-0416-220-2740
E-mail address: don.sasikumar@vit.ac.in

1877-0509 © 2020 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2020.03.228

inspiration is higher than it is during expiration. In abnormal breathing, the sounds (wheezes and crackles) generate certain pathological conditions of the airways on the lungs [2]. Wheezes are adventitious continuous sounds that have a wide frequency range. According to American Thoracic Society [3], a wheeze has a pitch of less than 400 Hz, and a duration of 250 msec. The latest study conducted by this same society concluded that the duration can be greater than 250 msec. The predominant peak value also varies according to age. For example, in the case of infants it is greater than 225 Hz, and in the case of adults it is above 400 Hz. Crackles are discontinuous sounds heard during inspiration having a duration of less than 100 ms. Based on the duration, crackles can be further classified as fine or coarse. Fine crackles are produced within the small airways, whereas coarse crackles arise from the bronchial segments. In the case of fine crackles, the duration is less than 10 msec. Coarse crackles have a duration of more than 10 msec. Respiratory sounds shows different acoustic properties depending on the subjects characteristics. Ana Oliveira [11] attempted a survey on detecting/characterizing respiratory sounds using computerized analyses. The respiratory sounds of 964 subjects were considered for the study. A systematic review of various classification techniques for lung sounds can be found in [4]. Feature selection is an important stage in lung sound classification. The Hjorth descriptor is mentioned as an excellent feature for lung sound classification in [5]. The article in [6] provides a thorough review on types of sensors, datasets, features, and the classification techniques used by the different researchers.

In recent years, machine learning techniques have proven their efficiency in medical applications. The advantages of applying machine learning techniques to lung sound analysis can be found in [7]. The automatic wheeze detection system is important in clinical decision-making. Even though commercial products are available in the market [8, 9], it is still under research. A brief discussion about the work carried out from 1985 to 2012 was presented in [10]. The survey was conducted on 27 articles on automated wheeze detection systems. These devices should be lightweight, portable, and easy to operate by people of all ages. Zhang [12] proposed a method for the automatic recognition of crackles based on spectral information. The dataset considered for the experiment consists of 50 normal sounds and 50 crackle events. The classification achieved an accuracy of 86% sensitivity with a specificity of 92%. Classifying three types of lung sound using statistical parameters of cepstral features was studied in [13]. The results demonstrate that segregating the sound into three subjects of study was promising. These cepstral features are effective in approximating certain information in classification, thereby minimizing the computational complexity. Convolution Neural Network architecture is a model that can be applied to a wide range of perceptual tasks. The CNN structure consists of a series of stages. CNN has been found to be effective in computer-vision-related applications. Murat Aykanat [14] tested the classification of respiratory sounds using CNN. In total, 17,930 lung sounds from 1630 subjects were considered. Experimental results show an accuracy of 80%. Chun Yu [15] developed a custom-made stethoscope, useful for monitoring the asthmatic condition of young children. The algorithm considered for the test was the adaptive respiratory spectrum correlation coefficient algorithm, which has high sensitivity, and low computational cost. The results show 88% sensitivity and 94% specificity in wheeze detection. Principle component analysis (PCA) has been widely used for exploring complex datasets. [16] proposed a method for asthma outcome prediction based on PCA and least square sum. It consists of three stages: PCA used for feature extraction and dimension reduction, classification achieved through least square sum, followed by performance evaluation. The performance achieved an accuracy of 95.54%. Like with any other biomedical signal, the signal complexities of lung sounds are evidenced by their fractal properties. It has been proven that fractal techniques are useful in the extraction of the characteristics of respiratory sounds. A comparison study of three fractal dimension algorithms was reported in [17]. The techniques were applied based on single variance, non-normalized signal morphology, and signal morphology normalized along both axes within the window. This study concludes that true changes in lung sound fractality can be measured using fractal dimensions. The amplitude and pattern characteristics of the lung sounds of children after bronchoconstriction is different from their base-line lung sound. [18] tested two different FD algorithms based on signal variance and morphology. The results show that a morphology-based fractal dimension offers a better classification ratio. Studies have shown that we can extract different features from lung sounds; we can generally group the feature selection into two approaches [19], i.e., classifier independent and classifier dependent. Classifier independent feature selection uses filter methods to identify the importance of features based on heuristic scoring criteria and the features with high scores are considered for the classification. The wrapper method is classifier dependent. Here, the subsets are selected based on the search mechanism, and the classifier is trained and tested to identify the appropriate feature subset. The steps are repeated until the specified requirements are met. The limitation of this approach is that the computational cost is high. Motivated by the success of statistical features in lung classification, a total of 15 different features taken

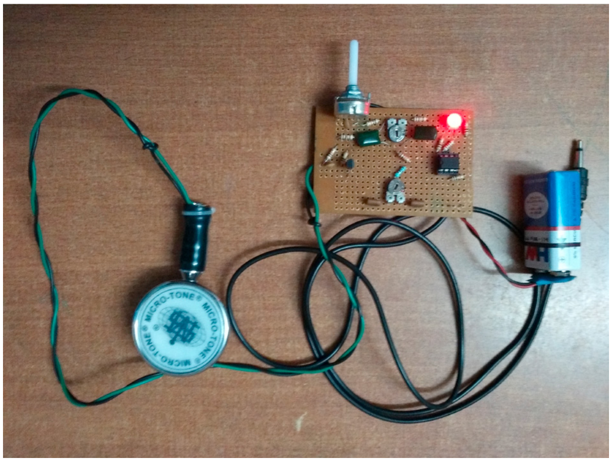

Fig. 1: Protoype Model

from lung sounds. In this study, a two class classification problem (normal and abnormal)is considered. Normal and abnormal lung sounds can be characterized based on duration, pitch, and sound quality. An expert can identify the sound heard using a stethoscope. Because the severity of breathing in asthmatic patients is high at night, it is ideally difficult for the examiner to examine all patients in a restrictive environment. To overcome this limitation, automated lung sound analysis is a potential choice. Several products are available in the open market Wheezometer, Wholter, STG etc. Wholter is used for home monitoring, whereby data are stored and uploaded to a system for further analysis. STG uses an electronic stethoscope with a P.C for further analysis. New products are available in the market with small sensors and are cost-effective for home monitoring. When examining an asthma patient, the observation is not purely based on the classification of lung sounds (normal, wheeze, and crackles). It also includes other prognostic factors including demographic, wheezing episodes, symptoms, parental history, housing condition, breathing test, pharmaceutical therapy, and asthma. Asthma in children has become common in recent years. Their symptoms develop before age five, and it is difficult to diagnose asthma at this age. Educating the parents to observe the childs wheezing frequency and wheezing apart from cold, and using other clinical history of the mother helps us gather common clinical parameters for the test. Feature selection is an important step in lung sound classification. We also conducted tests after reducing the dimension of the dataset by applying two different feature sub-selection methods. The first is random subset feature selection, and the second is sequential forward selection. By using these methods, we were able to identify a good feature subset from the available features. The data are tested using three different classifiers: SVM, Nave Bayes, and KNN. The remainder of this paper is structured as follows. Section 2 explains the various features and classification methods considered in this study. Section 3 explains the results and discussion. We summarize the work and mention some future directions in Section 4.

2. Materials and Methods

In this study, lung sound data were acquired from various online databases. These include the R.A.L.E repository [20], East Tennessee State University repository (ETSU) [21], Auscultation Skills Breath and Heart Sound (FAEMSE0 [22], Auscultation Assistant [23] and the data source provided in the CD from the book (Wilkins et.al [24]). We also developed a low-cost prototype to acquire lung sound recordings using a diaphragm and an omnidirectional condenser microphone inside a stethoscope shown in Fig.1. The acquired signals are then amplified and passed through two filters: the low-pass butterworth filter and the notch filter. The reduction in noise and the inherent aural clarity are important for extracting important information from the lung sound. Figs. (2,3,4) show sample Courserales, Crackle, and Wheeze signals that were considered. A total of 85 samples are considered for our study.To build a classifier, relevant statistical features from each cycle to train a classifier that can classify whether the lung sound belongs to the normal or abnormal class. Our goal in this work is to find a suitable subset of features from the selected features

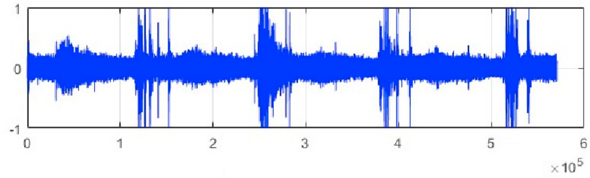

Fig. 2: Courserales Signal

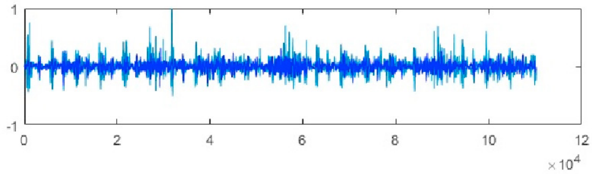

Fig. 3: Crackle Signal

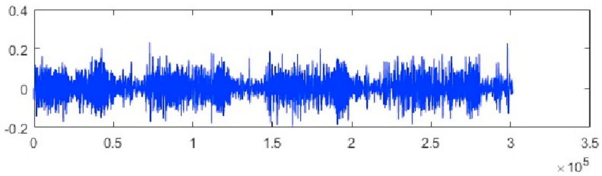

Fig. 4: Wheeze Signal

that best suits the classifier with an acceptable classification ratio. In our study, we used Higuchis and Katz fractal dimension, max and mean of spectral entropy, spectral flux, spectral rolloff,ZCR, max and mean of entropy and energy, and Kurtosis. Fifteen features are extracted from each cycle. These feature vectors can demonstrate considerable variation in their morphological structure to be identified as either a normal breath sound or an abnormal breath sound.

2.1. Higuchi 's and Katz's

Studies show that fractal dimension (FD) has proven to be an important feature in one dimensional signal classification. In our work, two different methods have been chosen for estimating the FD from lung sound signals: Higuchi's and Katz's. Higuchi 's method for estimating FD is efficient in discrete time series and less sensitive to noise.The algorithm calculates FD as follows. For a given time series x(1),x(2),...Ex(N), the algorithm constructs k new time series for m=1,2,...k.

$$L_k^m = \left[\frac{1}{k} \sum_{i=1}^{(N-m)/k} |x\,[m+ik] - x\,[m+(i-1)k)]| \,.\left\langle \frac{N-1}{(int((N-m)/k)*k)} \right\rangle \right] \tag{1}$$

where N is the total number of samples and , the normalized factor. For the time interval k, L(k) is called the mean of K values Lk for m = 1,2,3k. The data should fall on a straight line with a slope equal to FD[25]. Thus, Higuchis FD is the slope of the line that fits the pair ln[Lk], ln(1/k). Katzs method for estimating the FD depends on

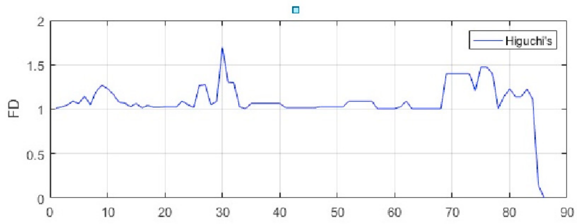

Fig. 5: Higuchi's FD

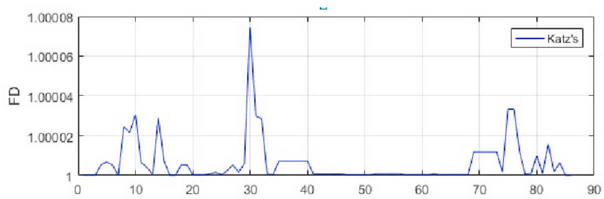

Fig. 6: Katz's FD

the morphological structure of the input pattern and is calculated as

$$KFD = \frac{log_{10(n)}}{log_{10(n)} + log_{10}(d/L)} \tag{2}$$

Where "n" is the number of increments between samples of the waveform over which FD is calculated, "d" is the value of the maximum distance measured from the beginning of the first increment, and "L" is the sum of all of the distances between successive increments. Fig.5 and Fig.6 shows the FD values obtained after applying Higuchi's and Katz's on a random set of samples considered for the experiment.

2.2. Entropy

Entropy is the measure of the amount of uncertainty associated with a system. It is commonly called Shannon entropy. Shannon entropy defines a measure H as a function of probability

$$H(p1, p2, ..., pn) = -\sum_{i=1}^{n} piln(pi) \tag{3}$$

2.3. Spectral Entropy

Spectral Entropy quantifies the spectral complexity of the lung sounds. It is obtained by first transforming the lung sound cycle to the Fourier spectrum. Once the power spectrum is calculated, the Shannon function is applied to map each frequency to a numeric value. Spectral entropy is the sum of all these values.

$$E = \sum_{n} P_n log\left(\frac{1}{P_n}\right) \tag{4}$$

2.4. Spectral Flux

Spectral Flux is the measure of how quickly the power spectrum of a signal is changing, calculated by the squared difference between the magnitudes of successive spectral distributions of the signal. It is written as

$$F_t = \sum_{n-1}^{N} (N_1(n) - N_{t-1}(n)^2) \tag{5}$$

2.5. Spectral Rolloff

Spectral Rolloff is the fraction of the bins in the spectrum at which 85% of the power is at lower frequencies. It measures the spectral shape and yields a higher value for a higher frequency. It is written as

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 \sum_{n=1}^{N} M_t[n] \tag{6}$$

2.6. Kurtosis

Kurtosis tells us whether the data is peaked or flat to a normal distribution. The kurtosis can be written as

$$K = \frac{\sum(X - \mu)}{\sigma} \tag{7}$$

2.7. Zero-Crossing Rate

Zero-Crossing Rate is the rate at which the signal changes from positive to negative or back. This feature is extensively used in speech recognition and is an important feature of percussive sound. It is written as

$$ZCR(m) = \sum_{m=-\alpha}^{\alpha} |sign[s(n)] - sign[s(n-1)]|\, w(n-m) \tag{8}$$

3. Feature Subset Selection

The selection of the most distinctive features for data classification is central to many pattern recognition problems. Finding the suitable set of features that best fit the data classes is a great challenge. One can arbitrarily select sets of features based on the type of classification problem they are addressing. Adding or removing features can either increase or decrease the performance of the classifier. If the numbers of feature sets is small, it can be processed faster. If the number of subsets of feature grown increasingly to meet the original feature size, then the best subset will be computationally expensive. At the same time, we also consider the over-fitting problem within the subset. In our work, the objective of feature subset selection is to reduce the number of features used to characterize the lung sound dataset to improve the classification accuracy. Several feature subset selection methods have been proposed by various researchers. Set covering problem is the most popular method for subset feature selection and is widely used in many application areas. In this work, we consider two existing methods for subset feature selection: Random Subset Feature Selection (RSFS) and Sequential Forward Selection (SFS) algorithms to choose the best subset of features for the classification of lung sounds. The subset is obtained by iteratively classifying the data with the K-NN classifier. In each iteration, the classifier arbitrarily selects subsets of all the features and adjusts the relevance of each feature according to the classification performance [26]. Unlike with other subset feature selection methods, where the relevance of a feature is computed by its presence or absence in the subset, RSFS selects a feature based on its average usefulness in the context of many other feature combinations. RSFS is less prone to convergence to a locally optimal solution. For the RSFS algorithm, the quality of the feature set improves as more iterations are performed. The algorithm for RSFS is given below.

Algorithm 1 Random Subset Feature Selection Algorithm

procedure RSFS
  for *each iteration in I in RSFS* **do**
    $S_i \leftarrow$ *Randomly picked subset of n features from full set F*
    $f_x(|S|) = n, x\epsilon[1, |F|]$
    $C_i \leftarrow$ *Criterion function value after performing KNN on given dataset using* S_i
    $E(C) \leftarrow$ *Expected value of* C_i
    Update there relevancies r_x
    $r_x \leftarrow r_x + C_i - E(c)$
  Repeat the step for other dataset

Algorithm 2 Sequential Forward Selection Algorithm

procedure SFS
  $FS \leftarrow$ Feature set,
  $S \leftarrow$ Required features
  for *i=0 to count(FS)* **do**
    for *j=i to count(FS)* **do**
      X=CalPerformance(FS[i])
      Y=CalPerformance(FS[j])
      if $Y \geq X$ **then**
        Swap i and j
  for *Each Feature in FS* **do**
    X=CalPerformance(F)
    Y=CalPerformance(S)
    Z=CalPerformance(F+S)
    if $X+Y \geq Z$ **then**
      $S=S \cup F$

Sequential forward selection (SFS) is a wrapper-based algorithm. It starts by creating an empty set. The best features are added to this empty set. These steps continue until the performance can no longer be improved [27].The general algorithm for SFS is given below. The average Mean and SD of the dataset prepared for the training is shown in Table 1.

4. Results and Discussion

In the classification stage, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naïve Bayes (NB) are used as classifiers. SVM is a supervised classification algorithm, where the classifier is used to find the decision surface that separates the data into two classes. The Nave Bayes classifier assumes a particular generative model for classifying the data. KNN is the most commonly used classification algorithm. It is an instance-based learning

Table 1: Average Mean ± SD of the dataset.

Features	Mean ± SD
Higichi's	0.114±0.143
Katz	1.000±1.142
Entropy Max	0.471±0.473
Entropy Mean	0.096±0.169
Flux Max	0.225±0.163
Flux Mean	0.033±0.019
Roll Off Max	0.086±0.066
Roll Off Mean	0.026±0.027
ZCR Max	0.053±0.004
ZCR Mean	0.018±0.018
Entropy Max	3.005±0.046
Entropy Mean	2.396±0.120
Energy Max	0.126±0.079
Energy Mean	0.015±0.014
Kurtosis	7.024±4.346

Table 2: The performance with all the features for classification with SVM, KNN, and Naïve Bayes.

Classifiers	Precision	Recall	F-Measure
SVM	93.8	83.3	88.2
KNN	86.7	72.2	78.8
Naïve Bayes	85.7	100	89.5

Table 3: The performance with SFS selected features for classification with SVM, KNN, and Naïve Bayes.

Classifiers	Precision	Recall	F-Measure
SVM	81.0	100	89.5
KNN	83.3	62.5	71.5
Naïve Bayes	76.2	100	86.5

procedure. This method considers the distances between features, which are calculated using Euclidean distance. This algorithm can work for both continuous and discrete features. The data is classified based on the majority vote of its neighbors. The classification accuracies are computed as Precision, Recall, and F-measure.

$$Precision(P) = \frac{TP}{TP + FP}, Recall(R) = \frac{TP}{TP + FN}, F - Measure = \frac{2.P.R}{P + R} \tag{9}$$

The dataset is split into training and testing set with a ratio of 60:40.Classification accuracy obtained by using five fold cross validation and the results are shown in Tables 2, 3 and 4. From these tables, one can observe that with a reduction in the number of features using RSFS and SFS, SVM, KNN, and Naïve Bayes show an improvement in F-measure for RSFS with (89.5%, 94.1%, and 89.5%). In the case of SFS, the SVM classifier shows an improved F-measure compared to KNN and Naïve Bayes (89.5%, 71.5%, and 86.5%). The original feature set has F-measures of 88.2%, 78.8%, and 89.5% for SVM, KNN, and Naïve Bayes.

Table 4: The performance with RSFS selected features for classification with SVM, KNN, and Naïve Bayes.

Classifiers	Precision	Recall	F-Measure
SVM	81.0	100	89.5
KNN	94.1	94.1	94.1
Naïve Bayes	81.0	100	89.5

5. Conclusion

In this paper, a study was done to demonstrate how the RSFS and SFS algorithms are capable of selecting important features for classifying lung sounds as normal or abnormal. RSFS provides significant improvement compared to SFS for three classifiers (SVM, KNN, and Naïve Bayes). The approach was tested on both offline data and real-time data acquired by our prototype and achieved F-Measure values of 89.5%, 94.1%, and 89.5% with features selected using RSFS. Future work in this area can focus on enhancing the acquisition system and enable the classification of various symptoms.

Acknowledgements

The authors thank VIT University for providing VIT SEED GRANT for carrying out this research work.

References

[1] Lin BS, Wu HD, Chen SJ (2015) "Automatic Wheezing Detection Based on Signal Processing of Spectrogram and Back-Propagation Neural Network",*J Healthc Eng*. **6** (4):649-72.
[2] Sarkar M, Madabhavi I, Niranjan N, Dogra M.(2015) " Auscultation of the respiratory system",*Ann Thorac Med*.**10** (3):158-68.
[3] Andrës E, Gass R, Charloux A, Brandt C, Hentzler A.(2018) " Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0.", *J Med Life*.**11**(2):89-106.
[4] Reichert S, Gass R, Brandt C, Andrs E.(2008)" Analysis of respiratory sounds: state of the art."*Clin Med Circ Respirat Pulm Med*. **2**:45-58.
[5] Achmad Rizal, Risanuri Hidayat and Hanung Adi Nugroho(2017)"Lung Sound Classification Using Empirical Mode Decomposition and the Hjorth Descriptor",*American Journal of Applied Sciences*. **14**(1):166-173.
[6] Renard Xaviero Adhi Pramono, Stuart Bowyer, Esther Rodriguez-Villegas(2017)"Automatic adventitious respiratory sound analysis: A systematic review",*PLoS ONE*. **12**(5).
[7] Rajkumar Palaniappan , Kenneth Sundaraj, Nizam Uddin Ahamed, (2013)"Machine learning in lung sound analysis: A systematic review",*Biocybernetics and Biomedical Engineering*.**33** (3):129-135.
[8] H.-K. Ra, A. Salekin, H.-J. Yoon et al.(2016) "AsthmaGuide: an asthma monitoring and advice ecosystem",*IEEE Wireless Health*:128135
[9] Ho-Kyeong Ra, Hee Jung Yoon, Sang Hyuk Son, John A. Stankovic, and JeongGil Ko,(2018) "HealthNode: Software Framework for Efficiently Designing and Developing Cloud-Based Healthcare Applications", *Mobile Information Systems*.**3** :1-12.
[10] Shaharum SM, Sundaraj K, Palaniappan R,(2012)"A survey on automated wheeze detection systems for asthmatic patients",*Bosn J Basic Med Sci*.**12**(4):249-55.
[11] Ana Oliveira and Alda Marques,(2014)"Respiratory sounds in healthy people: A systematic review",*Respiratory Medicine Volume*.**108** (4):550-570
[12] Zhang K,Wang X, Han F, Zhao H,(2015)"The detection of crackles based on mathematical morphology in spectrogram analysis",*Technol Health Care*.**23** (2):489-94.
[13] Nandini Sengupta ,Md Sahidullah and Goutam Saha,(2016)"Lung sound classification using cepstral-based statistical features",*Computers in Biology and Medicine*.**75**:118129
[14] Murat Aykanat, zkan Kl, Bahar Kurt and Sevgi Saryal,(2017)"Classification of lung sounds using convolutional neural networks",*EURASIP Journal on Image and Video Processing*,**65**:1-9
[15] Yu, Chun et al.(2013)"Soft stethoscope for detecting asthma wheeze in young children",*Sensors*,**13**(6): 7399-413.
[16] E.Chatzimichail, E. Paraskakis, M. Sitzimi, and A. Rigas,(2013)"An Intelligent System Approach for Asthma Prediction in Symptomatic Preschool Children",*Computational and Mathematical Methods in Medicine*.
[17] January Gnitecki , Zahra Moussavi,(2005)"The fractality of lung sounds: A comparison of three waveform fractal dimension alg",*Chaos, Solitons and Fractals*,**26**(4):10651072
[18] J. Gnitecki,Z. Moussavi, H. Pasterkamp,(2004)"Classification of Lung Sounds during Bronchial Provocation Using Waveform Fractal Dimensions",*IEEE EMBS*.
[19] Chuan Liu , Wenyong Wang , Qiang Zhao , Xiaoming Shen , Martin Konan ,(2017)"A new feature selection method based on a validity index of feature subset",*Pattern Recognition Letters*,**92**:18.

[20] Owens D. RALE,(2002) ”Lung Sounds 3.0.”,*Comput Inform Nurs*,**5**(3):910.
[21] Pulmonary Breath Sounds,East Tennessee State University repository(2002).
[22] Lippincott W.(2009)” Auscultation Skills: Breath and Heart Sounds”, the 3rd edition. Philadelphia, PA, USA.
[23] The Auscultation Assistant - Breath Sounds Datset
[24] Wilkins RL, Hodgkin JE, Lopez B.(2004)” Fundamentals of lung and heart sounds”, St. Louis, MO, USA: Mosby.
[25] Don.S, Duckwon Chung and Dugki Min,”Analysis of electrocardiogram signals of arrhythmia and ischemia using fractal and statistical features”,*Journal of Mechanics in Medicine and Biology*,**13**(1):1-15.
[26] Okko Räsänen and Jouni Pohjalainen,(2013)”Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech”,INTERSPEECH.
[27] A. W. Whitey, (1971)”A Direct Method of Nonparametric Measurement Selection”, *IEEE Transactions on Computers*,**C-20**,(9):1100-1103.