A Study of Accuracy of Data Mining Algorithms in Diagnosis of Emphysema Disease (EmD)

Abd Abrahim Mosslah

University of Anbar - College of Islamic

alaesawy24@yahoo.com

Received: 31 August 2017 Accepted: 13 September 2017

Abstract

Data mining techniques are the Amounts of actual data are used to analyze these data to predict whole some data to support a decision-making in a problem-solving. A data mining is very useful to analyze any disease characteristics to support the decision process and specify what the disease is and what Details IS. In the proposed present papers, using the real algorithms of data mining techniques to support different healthcare fields and adopted a correct decision about the diagnosis of emphysema disease and specify the risk factors for this disease to support decision process. In this research, a data-mining model of EmD prediction using a hybrid model Radial Basis Function - Neural Network (RBF-NN) and Genetic Algorithms (GA) has been presented. From the results, it has been seen that a hybrid model predicts EmD with nearly 95% accuracy. Furthermore, the examined samples of individuals share the same risk factors a symptom. Data mining depends on these symptoms and factors to diagnosis obstructive emphysema disease.

Keywords: Data Mining, Radial Basis Function Neural Network (RBF-NN), Decision Support (DS), Emphysema Disease (EmD), Genetic Algorithm (GA).

A Study of Accuracy of Data Mining Algorithms in Diagnosis of Emphysema Disease (EmD)

Abd Abrahim Mosslah

دراسة حول خوارزميات التنقيب عن البيانات للكشف عن مرض انتفاخ الرئة المزمن (Emphysema)

عبد إبراهيم مصلح

جامعة الانبار - كلية العلوم الاسلامية

الخلاصة

تقنيات استخراج البيانات هي طريقة تستخدم كمية من البيانات الفعلية لغرض تحليل هذه البيانات للتنبؤ بالبيانات لغرض دعم عملية اتخاذ القرار في حل المشكلة. ان عملية استخراج البيانات هي عملية مفيدة جدا لتحليل أي عملية خصائص لأي مرض لدعم عملية اتخاذ القرار وتحديد ما هو هذا المرض وما هي تفاصيله. في هذه الورقة البحثية المقترحة، قمنا باستخدام خوارزميات حقيقية لتقنيات استخراج البيانات لغرض دعم مجالات الرعاية الصحية المختلفة، واعتمد القرار الصحيح حول تشخيص مرض انتفاخ الرئة وتحديد عوامل الخطر من هذا المرض لدعم عملية اتخاذ القرار. في هذا البحث، تم بناء موديل مقترح للتنبؤ بمرض Emphysema Disease استخدام طريقة هجينة، هذه الطريقة الهجينة تتألف من الشبكات العصبية (RBF NN) والخوارزميات الجينية (GA). من خلال النتائج، فقد رأينا أن النموذج الهجين يتوقع Emphysema Disease مع ما يقرب من 95٪ من الدقة. وعلاوة على ذلك، فإن عينات فحص الأفراد تشترك مع نفس عوامل الخطر والأعراض. استخراج البيانات تعتمد على هذه الأعراض وعوامل لمرض انتفاخ الرئة وكذلك تشخيص الانسداد.

الكلمات المفتاحية: استخراج البيانات، الشبكة العصبية نوع دالة الاساس الشعاعية (RBF NN)، دعم اتخاذ القرار (DS)، مرض انتفاخ الرئة (EMD)، الخوارزميات الجينية (GA).

Introduction

Nowadays, Emphysema Disease is the major cause of death in the world. The World Health Organization (WHO) presents a report confirm that 14 million cases of death occur around the world because of EmD. The rate of death has increased over 80% by increasing the causes (family history, smoking, high blood pressure and other popular medical reasons) [1] [8]. The early, accurate prediction of the disease should reduce the risks of Emd. The diagnosis starts with symptoms depend on multi X-Ray of the lung to diagnose the disease exactly and which level became. The complete healthcare contains a huge data about any disease, especially Emd. This healthcare consists of concealed data, this data is very helpful and beneficial to making an

accurate making effective decision particularly Em patients. There is a huge amount of complex data related to a complete Healthcare belong to hospitals sources, disease diagnosis, a complete patient's computer records, medical treatments and etc., this huge data is the main key to analyze the total amount of data and then extract a knowledge in order to make an effective decision. In order to make a better accuracy of the diagnosis and minimize the diagnosis delay time, a system developed by using mixed techniques, artificial neural networks, and genetic algorithm in order to get efficient and reliable decision-making [2].

Theoretical Background

Most of the research work in disease detection is done towards enhancing and increasing the lifetime of the patient by proposing new ways of efficient healthcare:

In 2015, Alvaro D. Orjuela and Alexander S. Sepulveda built a new tool to respiratory disease discrimination based on acoustic lung signals and neural networks. This system deals with acquired signals by the lung and processed this signal by Mel Frequency Cepstral Coefficient to obtain representative parameters for neural network training. Results show that the use of ANN maintains the results for classification despite the differences between channels. At the same time, the classification rate drops 10% as a maximum, when these channel effects were analyzed, compared with no channel distortion [3].

In 2015, Almir Badnjevic, Mario Cifrek, Dragan Koruga and Dinko Osmankovic present a system for classification of asthma and Chronic Obstructive Pulmonary Disease (COPD) based on fuzzy rules and the trained neural network. The trained neural network depends on one thousand medical reports obtained from a database of the company Care Fusion were used. Afterward, the system was validated on 455 patients by physicians from the Clinical Centre University of Sarajevo [4].

In 2015, D. vally and CH. V. Sarma implement the back-propagation algorithm and genetic algorithm to compare the output accuracy percent for medical diagnosis on various chest diseases (Asthma, tuberculosis, lung cancer, pneumonia) [5].

In 2014, Ashok Mondal, Parthasarathi Bhattacharya, and Goutam Saha present a new method to distinguish between the normal and the abnormal subjects using the morphological

complexities of the lung sound signals. The performance of the proposed method is compared with a wavelet analysis based method. The developed algorithm gives a better accuracy of 92.86% and sensitivity of 86.30% and specificity of 86.90% for a composite feature vector of four morphological indices [6]. About detection of lungs status using morphological complexities of respiratory sounds.

System Methods and Materials

Medical diagnosis process begins when the patient with the doctor and offers a range of symptoms. Then the doctor more information asks the patient to identify symptoms in severe cases. It includes data collected from the patient's previous state of health, living conditions and other medical conditions. Physical examination of the patient's condition is completed, and in most cases, are conducted medical monitoring as well as the medical examination (s) of the patient before medical treatment. Imagery ((m) images taken by X-ray library has been ranked and classified to be suitable for our system.

1. Data Mining based model

Data mining is a traditional way to integrate the process of analyzing data with complex algorithms to extract accurate information, useful from among the huge amount of data is used; you may use those data later in the expectation an event in the future. Data is usually divided into three sections: (1) Data in the raw image without any examination or analysis is called Data. (2) Data has been analyzed and draw some simple information, which is called Information. (3) Information has been drawn in ways more complex, and if they were added to the analyst, the experience becomes our possession so-called Knowledge [7].

1`.1 Stages of the data mining

1. Initial analysis and conversion - preprocessing and transformation: are in that stage different raw data is converted to the form and to the standard format in the subsequent phase of analysis, the longer this stage is the most time-consuming and effort.

2. The liquidation and Assessment - Evaluation and filtration: This phase includes the emphasis on the validity and usefulness of the findings in order to be integrated into the decision-making process [14].

1.2 Algorithms for Data Mining

Data mining algorithms (and called machine learning) is a set of calculations lead to generate a data model. In data mining, the algorithm analyzes the acquired data and seeking for specific patterns to create a required model. The data mining algorithms can be classified into the following categories:

1. Classification: used to determine the compatibility of the new hypotheses of any of the categories defined in advance.
2. Regression: Measure the correlation between the different variables; it predict the value of a given variable based on another variable value.
3. Clustering: Those seeking the way to categorize the information to similar blocks in the characteristics of each block thus is similar in their characteristics from the other block.
4. Sequential Pattern Discovery: is a recurring pattern discovery in a certain order as buying medicines to cold followed by certain foods help to reduce the disease [9].

2. Disease Description

2.1 Emphysema Disease (EmD)

EmD is a disease of the lungs gradually over the long term, occurs when the destruction of the walls between the alveoli (alveolar), along with capillary blood vessels that operate within it. This reduces the total area within the lung where it meets the blood and the air together, limiting the possibility of the exchange of oxygen and carbon dioxide. Early in the disease, emphysema, be linked to inflammation in the small airways or bronchioles air, limiting the amount of air that can flow into the alveoli. The onset of symptoms in emphysema disease is usually after the passage of 50 years of age. The causes of EmD is: (1) Produces genetically due to the lack of Alpha-1 Antitrypsin (Alpha-1 Antitrypsin is a protein that protects the elastic structures in the

lungs to prevent adhesion) (2) Prolonged factors irritating the lungs of exposure, including: (a) Cigarette smoke (b) Air pollutants (c) Industrial fumes (d) Work in the coal mines [10].

2.2 Symptoms of EmD

The onset of symptoms in emphysema disease is usually after the passage of 50 years of age. Signs and symptoms of Emd appear as follows:

- **a.** Shortness of breath
- **b.** Uncomfortable
- **c.** Inability to exercise or daily activities easily
- **d.** Unjustified weight loss
- **e.** Chronic cough
- **f.** Rapid breathing.

2.3 The diagnosis of EmD

The doctor may resort to the following measures for the diagnosis of pulmonary emphysema as shown in (fig. 1):

- **a.** Clinical examination.
- **b.** Chest X-rays (X-ray).
- **c.** Computerized tomography (CT).
- **d.** Examination of pulmonary function.
- **e.** Laboratory testing of arterial blood gasses and measuring the concentration of gasses in the blood.
- **f.** Being a doctor examination to measure the anti-worker ratio Alpha-1 Antitrypsin-Alfa-1[11].

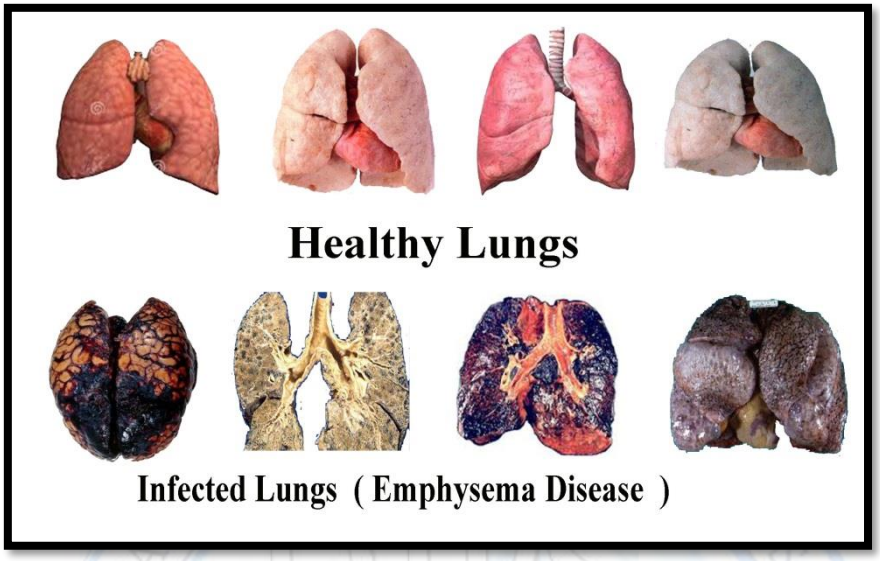

Figure 1: Healthy Lung Forms and Infected Lung Forms caused by Emphysema Disease

3. RBF Neural Network

The proposed robust system consists of (RBF) network as shown in (fig.2), (MLP) network and data mining. One of the novelist methods to be used with neural networks involves the hybridization of two types of techniques. The scientific truth behind the hybrid network is to succession phases of identical or not identical network models of examples to function as building blocks in analyzing a complex data structure. The hybrid proposed program developed by using RBF networks to minimize the scope of problem analysis in (MLP) network. Notice that the function of (RBF) network is collecting input features by characteristics of features distinguished. The (MLP) network has arranged those features categories into the entities desired to be recognized. To improve the behavior of networks, compare the work of each net individually. This artificial intelligent computer refers to a network size by the number of weights required for each (RBF) and (MLP) network. These results show that the hybrid approach can have some beneficial results. In addition, each stage of the network can be modified with a different paradigm to yield the desired results. Other benefits to using the hybrid techniques are to improve convergence speed and reliability of nets and fix the error cases on the global minimum by developing the structure of the input form to RBF network and

(MLP) network. The disadvantage with hybrid networks is suffered data accuracy lost [15]. To exceed this problem, Reduce the amount of input data in the first stage and pass only the input with robust features. This is leading to improve system speed, accuracy, and efficiency. When comparing (RBF) with other kinds of artificial neural networks, note that (RBF) network achieves very high accuracy for most of the data sets.

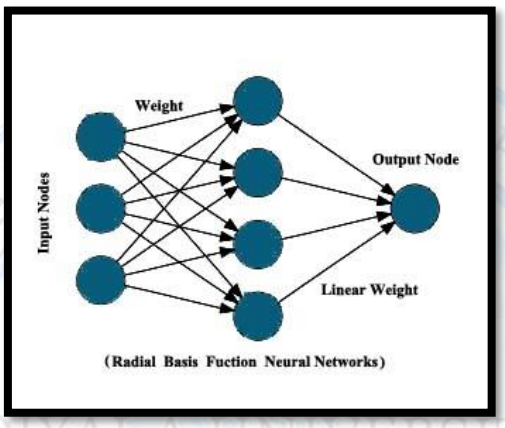

Figure 2: General Structure of RBF NN.

The proposed algorithm is based on historical databases of patients with (EmD), where this algorithm relies on the development of decision support for the prediction of (Em) [13]. In order to obtain satisfactory and accurate results, the algorithm depends on main medical parameters in order to accurately diagnosing the disease. The proposed ANN consists of an input layer to receive an input signal of (Em) case and output layer to produce the desired output. The input layer consisted of three neurons, the hidden layer had 25 neurons and the output layer had 5 neurons. The input layers (input values) are concerned with patient information (name, age, and sex). The middle layers called hidden layer (hidden values), these values are concerned with uses the patient data to make diagnoses. The output layer is called (output values), this layer produces the outputs (the values processing). The ANN learning start by putting an image as blocks, each block (iteration) consists of 8×8 matrix elements (PE) by multi iterations to suggested net [16]. The first iteration puts into the net as an input block and applying feedforward NN. First input iteration compares with desired output, if its, if there is an error

(defects), then adjust the weights of each node by applying back propagation NN for the same iteration until arriving at same desired output (i.e., during the training process these weights are changed in accordance with accomplish ideal exactness and coverage), at this time picking this iteration as a good solution [17].

4. System Based on GA

GA is an algorithm for optimization and machine learning based loosely on several features of biological evolution. They require five components:

a. A method for encoding answers for the issue on chromosomes.

b. An evaluation function, which returns a rating for each chromosome given to it.

c. A way of initializing the population of chromosomes.

d. Operators that might be connected to parents when they imitate to adjust their genetic composition.

e. Parameter settings for the algorithm, the operators, and so on [18].

Given these five components, a (GA) operates according to the following two steps:

1. Initialize the population using the initialization procedure, and evaluate each member of the initial population.

2. Reproduce until a stopping criterion is met. Reproduction comprises of iterations of the accompanying three stages:

a. Choose one or more parents to reproduce. Selection is stochastic; however, the individuals with the highest evaluations are favored in the selection.

b. Choose a genetic operator and apply it to the parents.

c. Evaluate the children and accumulate them into a generation. After accumulating enough individuals, insert them into the population, replacing the worst current members of the population [19] [20].

5. Hybrid Proposed Algorithm Based Detection Em System: A Complete Description

The proposed algorithm is based on historical databases of patients with Em, where this algorithm relies on the development of decision support for the prediction of Em. This algorithm depends on main medical parameters such as: (shortness of breath, uncomfortable, inability to exercise or daily activities easily, unjustified weight loss, chronic cough, and rapid breathing). The proposed RBF NN consists of an input layer to receive an input signal of Em case and output layer to produce the desired output to be an entry to the GA. The input layer consisted of three neurons, the hidden layer had 25 neurons and the output layer had five neurons. The input layers (input values) are concerned with patient information (name, age, and sex). The middle layers called hidden layer (hidden values), these values are concerned with uses the patient data to make diagnoses. The output layer is called (output values), this layer produces the outputs (the values processing). The RBF NN learning starts by putting an image as blocks, each block (iteration) consists of 8×8 matrix elements (PE) by multi iterations to suggest net. The first iteration puts into the net as an input block and applying feedforward NN. First input iteration compares with desired output, if it's, if there is an error (defects), then adjust the weights of each node by applying back propagation NN for the same iteration until arriving to same desired output, at this time can pick this iteration as a good population according to the map described in (fig.3).

In the GA, the thresholding is a method to the process of natural selection the beginning, a small number of chromosomes may survive. The new population would be generated, to find some chromosomes that pass the test. In order to produce parents' offspring, the threshold allows some of the chromosomes to continue when the threshold has value more than the cost value of chromosomes. Most of the chromosomes will survive provided the threshold is not changed in the next generations.

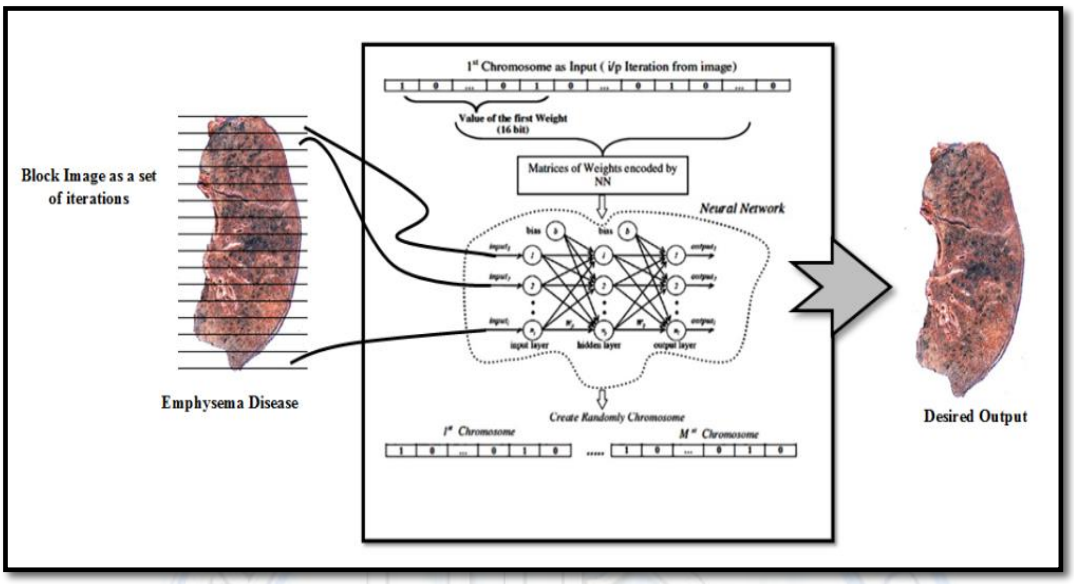

Figure 3: Sample Iteration Learned by NN_GA Algorithm

The learning begins with RBF neural network, by putting an Em images as a sets blocks called (iterations). Each iteration consists of 8×8 matrix elements (process elements) by multi iterations to the net. The first iteration enters the network as a set of t block and applies feed-forward neural network. The first iteration is input and compares with the desired output, if the both are equal; this means obtained the appropriate output. If there are defects (error), then adjust the weights of each processing element (node) by applying feedback neural network for the same block (iteration) (i.e. during the training process these weights are adjusted to achieve optimal accuracy and coverage). After learning with both (MLP) and (RBF) networks, the output of the learned network iterations heading towards genetic algorithms (GA) directly to begin the training phase to choose the best chromosome. GA is dealing with the inputs from (RBF) and (MLP) networks as a chromosome. Finally, the best chromosome has good fitness is representing the best solution. MLP network was used for image classification. The (RBF) network was used for neural network training. According to the evolutionary algorithm, a genetic algorithm starts with a population (collection) of individuals, which evolves toward optimum solutions through the genetic operators (selection, crossover, mutation), inspired by biological processes. Each element of the population is called chromosome and codifies a point

from the search space. The search is guided by a fitness function meant to evaluate the quality of each individual. The efficiency of a genetic algorithm is connected to the ability to define good fitness function. The optimization will involve the random searching for the optimal values of the weights assigned to the connections between the neurons within the network where each processing element represents a neural network with a particular set of weights. The aim of the hybrid algorithm is to find the population producing the smallest value of the error function. Where is the chart, the steps below represent a complete description on of unsupervised neural networks and genetic algorithm for the proposed system:

10 Input (x1,x2,x3,…xn) /image parameters

Start Neural Network Training

Compare Input parameter with Em image parameter/ desired out

If equal (Matching) then 30

Net. Weight adjusted: Goto 10

30 Start GA Training

40 Applied GA Operators

Calculate the error in each neuron

If tolerance <> Convergence then 40

Best Solution

Note that, the input parameters of RBF NN are shortness of breath, uncomfortable, inability to exercise or daily activities easily, unjustified weight loss, chronic cough, and rapid breathing.

Experimental Results

Tables below of proposed model are organized as follows: Table 1 and 2, experimental results of training parameters acquired by RBF-NN and GA each individually. Table (3), show experimental result by applying NN-GA System.

Table 1: Results acquired by (RBF-NN) training parameters

No. of EmD Images	Successful Matching	Unsuccessful Matching	Training Time by second	Image Efficiency (in percentage)
6	4	2	0.28	55%
14	11	3	0.36	80.8%
20	17	3	0.39	72%
22	16	6	0.47	79.55%
25	19	6	0.51	81.01%

Table 2: Results acquired by (GA) training parameters

No. of Em Images	Successful Matching	Unsuccessful Matching	Training Time by second	Image Efficiency (in percentage)
6	5	1	0.24	59 %
14	11	3	0.32	71.05 %
20	14	6	0.47	72.2 %
22	16	6	0.51	83.1 %
25	19	6	0.52	82.05 %

Table 3: Experimental result by NN_GA (Hybrid Model) System

No. of Em Images	Hybrid Model Matching	Error Rate	Time Training (in second)	Mean Accuracy (%)	RMSE
6	5	2	0.1	81%	0.3
14	13	2	0.31	79%	0.23
20	18	3	0.21	82%	0.45
22	20	4	0.4	81%	0.35
25	23	6	0.5	95%	0.46

Where RMSE: Root Mean Square error

Note that, increasing the number of training iterations leads to consumption of effort and time. On the other hand, choose a few number of training of iterations could lead to non-network training, and thus its inability to discrimination. Increase the number of training in neural networks leads to reduce the Error Rate (ER) down to the desired output. If the network did not reach to the desired output, the process of weights adjustment starts at this stage with each iteration. The process of weights adjustment includes the feed forward and back propagation down to the best values of the output, so weights adjustment includes all the nodes in the neural network. Now, during the training, the network continuously weights adjustment that associated with each node in order to output be closer than desired output and this is based on Learning Rate (LR). One of the main reasons for Non-neural network learning is not to choose

appropriate ratio LR. LR is used to control the rate of increase or decrease the value of the weight values in the learning stage. The training stage is continuing with all nodes inputs and weights adjustment related to each node in the network. The output value is closer to desired output. After reviewing the results through training of neural networks, genetic algorithms and hybrid system, note that training results through hybrid systems are better than the other systems due to hybrid system used properties of neural network and genetic algorithms to producing a new system holds all the properties of the artificial intelligence systems and it has been demonstrating through above results. When comparing the results with previous experience in the same field using the genetic algorithm to determine the best number of learning sets incoming from neural networks, which have best weight aggregates, which makes the network more efficiently training through hybridization.

Conclusion

In this research, a system developed for predicting EmD of a patient. RBF-NN and GA have been applied in different areas of medical research due to provide more results that are accurate when it's used in medical tests and diagnosis. The prediction is done based on historical Em database. The RBF-NN and GAs are capable of learning through examples and to generalize with the power of pattern recognition and distinction tasks. They are mathematical models used for understanding and predicting complex and chaotic dynamics in complex biological systems. The system is developed by using ANN technology and GA for the prediction of Em with high accuracy. To achieve high accuracy, the system builds by the mixed genetic algorithm will be involved with neural network technology.

References

1. A.A, "World Health Organization", Global Status Report on Non Communicable Diseases pp.1-176, 2011.
2. D.T. et al, "Disease Control Priorities in Developing Countries", Second Edition, World Bank, 2006.
3. A. D. and A.S., "Respiratory Disease Discrimination Based on Aacoustic Lung Signals and Neural Networks", Signal Processing, Images and Computer Vision (STSIVA), 20th Symposium on., 2015.
4. A.B. et al., "Neuro-fuzzy Classification of Asthma and Chronic Obstructive Pulmonary Disease", BMC Medical Informatics and Decision Making, 2015.
5. D. V. and CH. V., "Diagnosis Chest Diseases Using Neural Network and Genetic Hybrid Algorithm", Journal of Engineering Research and Applications(JERA) , Vol. 5, Issue 1(Part 2), p.p.20-26, Jan., 2015.
6. A.M. et al. ,"Detection of Lungs Status Using Morphological Complexities of Respiratory Sounds" , the Scientific World Journal,Vol.,2014.
7. D. S. et al.," Decision Support System for Medical Diagnosis Using Data Mining", International Journal of Computer Science(IJCS), Issues, Vol. 8, Issue 3, No. 1, May, 2011.
8. Global Burden of Disease, 2004, World Health Organization, update 2008.
9. H. et al., “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.
10. A.M,. et al.," Classification of Respiratory Abnormalities Using Adaptive Neuro-Fuzzy Inference System",. Int. Inf. and Database Sys. Lecture Notes in Computer Science, 2012.
11. H.N., "An integrated software package for model-based neuro-fuzzy classification of small airway function", ETD Collection for University of Texas, 2009.
12. L.O. and P.A., " Hybrid Training of Radial Basis Function Networks in A Partitioning Context of Classification", Neurocomputing, Vol. 28. Nos. 1-3 p.p. 165-175, 2001.
13. M.N. et al. ,"Decision Support System for Asthma", International Journal of Information and Computation Technology, 3:p.p. 549-554, 2013.

14. A.S. et al., " Design and Implementation of a Fuzzy Expert System for Detecting and Estimating the Level of Asthma and Chronic Obstructive Pulmonary Disease", Middle-East Journal of Scientific Research, 14:1435-1444, 2013 .

15. B., A. et al., "Clustering Algorithms for Radial Basis Function Neural Network", p.p. 2320-8945. 2013.

16. B.,J.A." Radial Basis Function Networks: Introduction", 2004.

17. S., A. et al. ,"A new algorithm for developing dynamic radial basis function neural network models based on genetic algorithms", Computers and Chemical Engineering, 28, p.p. 209-217, 2003.

18. D.E. “Search in GA , Machine Learning Optimization”, Singapore, 2002.

19. M. F.,Ch. K., “Industrial Applications of Genetic Algorithms”. Boca Raton, FL: CRC Press, 1999.

20. A. Shawkat," Comparison between some Data Mining Algorithms in Diagnosis of Gastric Ulcer", The 5th International Scientific Conference, The Union of Arab Statistician, Egypt, 2016.