Diagnosis Lung Cancer Disease Using Machine Learning Techniques

Shokhan M. Al-Barzinji

College of Computer Science and Information Technology, University of Anbar, Iraq

shokhan_albarzinji@yahoo.com

Abstract

Lung cancer (LC) is the leading cause of cancer-related deaths, both in women and among men. Yearly, LC kills more people than other cancers such as colon cancer, prostate cancer, and lymphoma and breast cancer, with 2.8 million deaths in 2017. To analyze any disease characteristics, a data mining is used for decision support process, specify the disease with its details. Data mining techniques are the amount of actual data are used to analyze these data to predict wholesome data to support a decision-making in a problem-solving. In this paper, used a data mining techniques, hybrid model Radial Basis Function - Neural Network (RBF-NN) and Genetic Algorithms (GA) to support different healthcare fields and adopted a correct decision about the diagnosis of LC disease and specify the risk factors for this disease to support decision process. The results demonstrate that the prediction accuracy of LC through the hybrid method is about 94%.

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

المستخلص

يعد سرطان الرئة المسبب الرئيسي للوفيات لاغلب مرضى السرطان ، وفي كلا الجنسين. اذ يفتك سرطان الرئة بعدد كبيرا من الأشخاص مقارنة مع الأنواع الأخرى من السرطانات مثل سرطان القولون وسرطان البروستاتا والورم الليمفاوي وسرطان الثدي حيث وصلت عدد الوفيات عام 2016 فقط الى 2.6 مليون حالة وفاة. ولغرض تحليل أي مرض، يتم استخدام استخراج البيانات لدعم عملية اتخاذ القرار وتحديد المرض وتفاصيله. ان تقنيات استخراج البيانات هي كمية البيانات المستخدمة لتحليل عملية صنع القرار في حل المشكلات. في هذه الورقة البحثية ، تم استخدم تقنيات استخراج البيانات ، وكذلك نموذج هجين مكون من – الشبكات العصبية (RBF-NN) والخوارزميات الجينية (GA) لغرض دعم مختلف مجالات الرعاية الصحية واعتمد القرار الصحيح بشأن تشخيص مرض سرطان الرئة وتحديد عوامل الخطر لهذا المرض لدعم عملية اتخاذ القرار. تُظهر النتائج أن دقة التنبؤ بمرض سرطان الرئة من خلال الطريقة الهجينة تبلغ حوالي 94٪.

1. Introduction

Recently, cancer disease is the main cause of death in the world. According to WHO report (World Health Organization), 15 million cases of death occur around the world because of cancer. The average 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 prediction of the disease should decrease the risks of cancer. The diagnosis starts with symptoms based on multi X-Ray of the lung to diagnose the accurate disease. The complete healthcare contains a huge data about any disease, such as cancer. This data is very helpful to making an accurate effective decision particularly cancer patients. There are a big compound data related to a healthcare belong to hospitals sources, disease diagnosis, a complete patient's computer records, medical treatments and etc., this data is the major key to analyze the total amount of data and then extract a knowledge in order to make an effective decision. In order to 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].

2. Related Work

Most research work on disease detection is done to promote and increase the patient's life by proposing new ways of effective healthcare:

1. In 2010, Zakaria Suliman and Reema Ashaibani present a several data min methods for Early Diagnosis of LC. Methods in this paper classify the digital X-ray chest films in two categories: normal and abnormal. The normal ones are those characterizing a healthy patient. The abnormal ones include Types of LC [1].
2. In 2012, Mokhled S. AL-TARAWNEH, provide a technique to improve the medical image where the technology depends on the image quality and accuracy, which is the main factor of

this research and evaluation of image quality as well as improvement depends on the stage of improvement where low pre-treatment techniques are used on the basis of Gabor filter within Gaussian rules. The proposed technique is effective for principles of segmentation to be a ROI foundation for obtaining the feature extraction. The proposed method gives a good outcome if comparing with other methods where based on general features [2].

3. In 2013, V. Krishnaiah, G. Narsimha and N.Subhash Chandra, present a method for early detection and correction diagnosis of the disease that helps the physicians in patient life-saving[3].
4. In 2014, Zakaria Suliman Zubia and Rema Asheibani Saad, purposed and enhancement an automatic method for early detection of LC. These system developed by analyzing X-ray chest images using varied steps [4].
5. In 2015, T.Christopher and J.Jamera present a study of data mining techniques used to detect early stage of LC disease using data mining algorithm such as data mining, LC prediction, classification, decision table, naïve based, and ant colony optimization [5].
6. In 2016, T.Christopher and J.Jamera present an analyzed the LC prediction model by using classification algorithm such as Naive Bayes, Bayesian network and J48 algorithm [6].

3. Materials and System Methods

Clinical medical treatment begins when the patient advances to the treating physician through a range of clinical symptoms. The specialist asks patient more queries to extract more information about the disease to specify more symptoms, especially in emergency cases. The data collected from queries includes patient's medical history, living conditions, and other medical parameters. The patient's condition physical screening is completed, and in most cases are conducted medical surveillance as well as the medical screening of the patient before medical treatment. The X-ray images have been classified to be suitable for the proposed method [7].

4. Disease Description

4.1 Lung Cancer (LC)

LC is simply uncontrolled growth of cancer cells and abnormal in one or both lungs. These blocks of these cells form cancerous tumors that impede healthy lung activity [12].

4.2 Symptoms of LC

Symptoms of LC in most cases do not appear in their initial stages. LC symptoms often occur only when the disease has already reached an advanced stage. LC symptoms include:

a. A cough appears and does not disappear.
b. Changes in a chronic cough or cough of smokers.
c. A cough accompanied by a blood clot called (Hemoptysis)

d. Tightness of breath.
e. Chest pains.
f. Bare sound.
g. Pains in the shoulder area and around the result of pressure on the nerves of the tumor (Pan cots Syndrome)

SVCS (Superior Vena Cava Syndrome): a feeling of fullness in the head and shortness of breath, an increase in veins in the chest [13]. Fig (1) below, shows the cancer stages in the human lung.

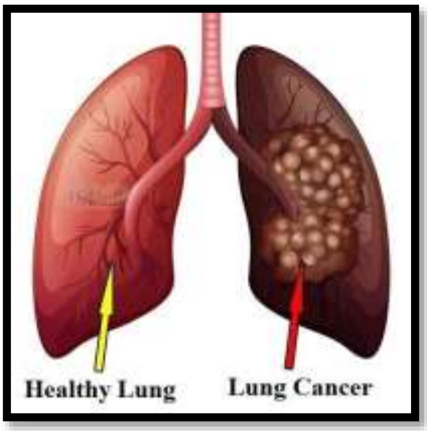

Fig. 1, Healthy and Non Healthy Lung

4.3 Diagnosis of LC

Doctors are still unsure whether screening tests should be performed to detect LC, or not. Even if the patient belongs to one of the risk groups most likely to develop LC, it is not clear whether X-ray or even computed tomography (CT) is useful. Research suggests that such tests can, in some cases, lead to early detection of the disease, at a relatively early stage during which cancer can be treated with great success. Recently, there have been recommendations that X-ray and computed tomography (CT) scans can be performed for smokers over the course of 15 or more years between the ages of 55 and 77 [14]. In order to diagnose LC, the doctor recommends different tests, including:

a. Imaging Tests
b. Saliva examination (examination of the shape of cells and their functions / cytology - Cytology)
c. Tissue Examination (Biopsy - Biopsy) [15]. Below fig.(2), the representative figure of the healthy lung and infected lung by cancer.

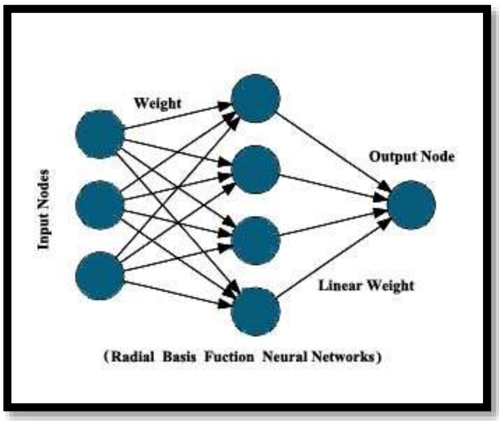

Fig. 2: General Structure of RBF NN.

5. Proposed Scheme based on Data mining

Data mining are conventional methods to integrate the process of analyzing data with complicated algorithms to extract delicate information, helpful between the huge amounts of data is used. The data can be classified into three categories: (1) Raw image data, without any examination or analysis, these data are called the main data. (2) Information, when data has been analyzed, draw some simple information and extract meaningful information. (3) Knowledge, these experience by representing information in different complicated methods, additionally added to the analyst [8].

5.1 Data Mining Stages

PAT (Primary Analysis and Transformation) - pre-processing and conversion: at this stage, the various raw data are converted to the model and to the standard form in the subsequent phase of analysis, and this phase is the most time-consuming and effortless.

FE (Filtering and Evaluation) - Evaluation and Nomination: this phase involves focusing on the validity and usefulness of the results for their integration into the decision-making process. [9].

5.2 Algorithms of Data Mining

The algorithms of data mining (another called machine learning) is a calculation set led to producing a data model. The algorithm in data mining (in machine learning) analyzes the gained data and investigate for particular patterns to generate the desired model. The data mining algorithms can be divided into a below various categories [10]:

a. Classification: used to define the compatibility of novel hypotheses for any predefined categories.

b. Regression: to measure the relationship between various parameters, they predict the value of a given variable based on another variable value...

c. Clustering: a way to classify information into identical blocks according to the properties of each block, and then its identical in their properties from another block.

d. Sequential Pattern Detection: Is the discovery of a recurrent pattern in a specific order as the buying of cold medicines followed by particular foods that assist to disease reducing [11].

6. RBF Neural Network

The proposed model consists of a network (RBF) as shown in (fig.2), and data extraction. One of the most innovative methods for use with neural networks involves the hybridization of two types of techniques. The scientific truth behind the hybrid network is the succession of identical or different network model phases to serve as building blocks in the analysis of a complex data structure. The proposed hybrid program was developed using RBF networks to minimize the scope of network problem analysis. Note that the network function (RBF) collects input characteristics by characteristics of distinguished features. These artificial intelligent computers have a network size by the number of weights required for (RBF) network. These results show that the hybrid approach can have beneficial results. In addition, each step of the network can be modified with a different paradigm to achieve the desired results. Other advantages of the use of hybrid techniques are to improve the speed of convergence and the reliability of the networks and correct the cases of error in the global minimum by developing the structure of the registration form in the RBF network. The disadvantage of hybrid networks is the loss of data accuracy [15]. To overcome this problem, reduce the amount of input data in the first step and pass only the input with robust features. This leads to improve the speed, precision, and efficiency of the system. When comparing (RBF) with other types of artificial neural networks, keep in mind that the network (RBF) achieves very high accuracy for most data sets.

The proposed algorithm is based on historical databases of patients with LC disease, where this algorithm is based on the development of decision support for the prediction of LC [13]. To obtain satisfactory and accurate results, the algorithm depends on the main medical parameters to diagnose the disease accurately. The proposed RNA consists of an input layer to receive an input signal from the packet LC and the 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) refer to the patient's information (name, age and sex). The intermediate layers called hidden layer (hidden values), these values are related to the use of patient data to make diagnoses. The output layer is called (output values), this layer produces the outputs (the processing of the values). The ANN training begins by placing an image in blocks, each block (iteration) consisting of (8×8) matrix elements (PE) by multiple iterations to the suggested network [16]. The first iteration places the network as an input block and applies NN feedforward. First iteration of input compared to the desired output, if, if there is an error (defects), adjust the weights of each node applying the inverse propagation NN for the same iteration until reaching the same desired result (that is, during the training process the weights are changed according to the ideal precision and coverage), choosing this iteration as a good solution [17].

7. System Based on GA

GA are algorithms for optimization and machine learning based on several characteristics of biological evolution. They require five components:

a. A method for encode answers for the issue about the chromosomes.

b. An evaluation function, which returns an evaluation for each chromosome assigned to it.

c. A way of initialize the population of chromosomes.

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

e. Parameter configuration of the algorithm, operators, etc. [18].

Given these five components, one (GA) operates in 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. The reproduction includes iterations of the accompanying three stages:
   a. Choose one or more parents to reproduce. The 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].

h. A Complete Description of a Hybrid Proposed Algorithm

The proposed algorithm is based on historical databases of patients with LC, where this algorithm is based on the development of decision support for LC prediction. This algorithm depends on key medical parameters, such as: (severe coughing, difficulty breathing, change in sound, pain in the chest, shoulder or back and change in color of esophageal secretions). The proposed NB RBF consists of an input layer to receive an input signal from the LC packet and the output layer to produce the desired output as input 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 values of input layers refer to the patient's information (name, age and sex). The intermediate layers called hidden layer (hidden values), these values are related to the use of patient data to make correct diagnoses. The output layer is called (output values), this layer produces the outputs (the processing of the values). The RBF NN training begins by placing an image in blocks, each block (iteration) consists of (8×8) matrix elements (PE) for several iterations to the suggested network. The first iteration places the network as an input block and applies NN feed forward. First iteration of input compared to the desired output, if there are errors then adjust the weights of each node applying NN inverse propagation for the same iteration until reaching the same desired output, at this moment you can choose this iteration as a good population according to the map described in (fig.3)

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

The learning with RBF neural network start by putting an images LC 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 comes in 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 (RBF) network, 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) network as a chromosome. Finally, the best chromosome has good fitness is representing the best solution. 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. In the below full chart, the infrastructure of the neural network, (GA) and image processing. In the below steps, a complete description on unsupervised neural networks and genetic algorithms algorithm of proposed system:

As a noted, the input parameters of RBF NN are as a noted, the input parameters of RBF NN are severe coughing, difficulty breathing, change in sound, pain in the chest, shoulder or back and change in color of esophageal secretions (phlegm).

The below tables of proposed method 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.

The increasing number of iterations of training leads to a consumption of effort and time. On the other hand, choosing some iteration trainings could lead to training outside the network and, therefore, to their inability to discriminate. Increasing the amount of training in neural networks leads to reduce the error rate (ER) to the desired output. If the network has not reached the desired result, the process of weight adjustment begins at this point in each iteration. The process of weight adjustment includes the propagation of the advance and the return to the best values of the output, so that the weight adjustment includes all the nodes in the neural network. Now, during training, the network continuously weighs the adjustment associated with each node to produce an output closer to the desired output and this is based on the learning rate (LR). One of the main reasons for non-neuronal network learning is not choosing the right LR ratio. LR used to control the rate of increase or decrease in the value of weight values in the learning phase. The training step continues with all the node entries and the weight adjustment related to each node in the network. The output value is closer to the desired output. After reviewing the results by training neural networks, genetic algorithms and the hybrid system, keep in mind that the training results for hybrid systems are better than those of other systems due to the hybrid properties of the neural network. And genetic algorithms. Artificial intelligence systems and demonstrated it through the previous results.

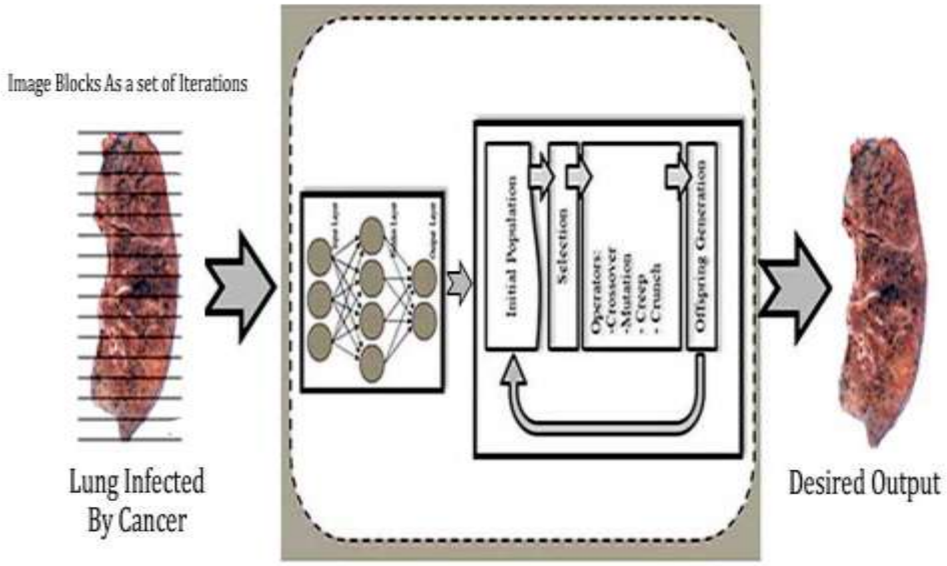

Fig. 3 Proposed Structure of Lung Cancer Detection using Hybrid Algorithm

To determine the best number of learning groups contained from neural networks, comparing the results with previous experience in the same field using the genetic algorithm which has the best weight totals, making the network more efficient training through hybridization.

9. Experimental Results

The Experimental results of proposed algorithm can be describe as below:

1. Input (x1, x2, x3, xn) /image parameters

Start Neural Network Training

2. Compare Input parameter with LC image parameter/ desired out

```
    If equal (Matching) then 3
    Net. Weight adjusted: Go to 1
3. Start GA Training
4. Applied GA Operators
    Calculate the error in each neuron
    If tolerance <> Convergence then 4
    Best Solution
```

In the 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 Artificial Neural Networks training parameters.

No. of LC	Successf	Unsucces	Training	Image
8	5	3	0.29	57%
16	12	4	0.37	81.8%
22	18	4	0.40	73%
24	17	7	0.48	80.55%
28	20	8	0.53	82.01%

Table 2. Results acquired by Training Parameters of Genetic Algorithm

No. of	Succes	Unsucces	Training Time by	Image
8	6	2	0.23	60 %
16	13	3	0.35	72.05 %
22	14	8	0.49	73.2 %
24	16	6	0.52	84.1 %
28	19	9	0.53	83.05 %

Table 3. Experimental result by Hybrid Model Algorithm

No. of	Hybrid	Error	Time	Mean	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.

10. Conclusion

A system developed for predicting LC disease of a patient produced in this paper. The prediction is done based on historical LC database. The ANN 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. Development of a system by using

ANN technology and GA for the prediction of LC with high accuracy. To achieve high accuracy, the system builds by the mixed genetic algorithm will be involved with neural network technology.

Reference

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