DOI: 10.21917/ijsc.2014.0115

A NOVEL SHAPE BASED FEATURE EXTRACTION TECHNIQUE FOR DIAGNOSIS OF LUNG DISEASES USING EVOLUTIONARY APPROACH

C. Bhuvaneswari[1], P. Aruna[2] and D. Loganathan[3]
[1, 2]Department of Computer Science and Engineering, Annamalai University, India
E-mail: [1]bhuvana.csdept@gmail.com, [2]arunapuvi@yahoo.co.in
[3]Department of Computer Science and Engineering, Pondicherry Engineering College, India
E-mail: drloganathan@pec.edu

Abstract

Lung diseases are one of the most common diseases that affect the human community worldwide. When the diseases are not diagnosed they may lead to serious problems and may even lead to transience. As an outcome to assist the medical community this study helps in detecting some of the lung diseases specifically bronchitis, pneumonia and normal lung images. In this paper, to detect the lung diseases feature extraction is done by the proposed shape based methods, feature selection through the genetics algorithm and the images are classified by the classifier such as MLP-NN, KNN, Bayes Net classifiers and their performances are listed and compared. The shape features are extracted and selected from the input CT images using the image processing techniques and fed to the classifier for categorization. A total of 300 lung CT images were used, out of which 240 are used for training and 60 images were used for testing. Experimental results show that MLP-NN has an accuracy of 86.75 % KNN Classifier has an accuracy of 85.2 % and Bayes net has an accuracy of 83.4% of classification accuracy. The sensitivity, specificity, F-measures, PPV values for the various classifiers are also computed. This concludes that the MLP-NN outperforms all other classifiers.

Keywords:

Feature Extraction, Multilayer Perceptron, Neural Networks, Bayes Net, Sensitivity, Specificity, F-Measure

1. INTRODUCTION

Lung diseases are one of the major health problems in the world that refers to disorders effect on the lungs. Inhalation prevents the body from getting sufficient oxygen. It can be caused by infection, an exposure at the workplace, medications and various disorders. Some of the lung diseases are emphysema, Asthma, chronic bronchitis Infections such as pneumonia, Lung cancer and Sarcoidosis. The causes of all types of lung disease, includes: smoking, radon, asbestos and air pollution. Lung disease has different signs and symptoms. some of the signs are cough, pain while breathing, trouble in breathing etc. imaging techniques like X-rays, CT scans provide images of tissue types like lung, bone, tissue etc., A computed tomography (CT) scan or computerized axial tomography which uses X- rays and computer to create detailed images of body's inner parts. For the diagnosis of the lung diseases computed tomography is used for investigation.

Bronchitis can be characterized by a bronchial wall damage which is associated with a significant increase in airway diameter. A patient with symptomatic bronchitis has chronic bacterial or fungal infection leading to inflammation of the airways. The manifestation of the disease is due to excessive mucus production causing impaired clearance of the same. The symptoms of bronchitis are expectoration of abnormal mucus, cough, airflow obstruction and respiratory tract infection. Pneumonia is inflammation of the parenchyma of the lung. Streptococcus pneumonia, are the most common cause of Pneumonia. Organisms settle in small air sacs called alveoli and continue multiplying and affect the body's immune system. Pneumonia are caused by white blood cells attacking the infection, the sacs get filed with fluid and pus.

In this work, a computerized approach for classification of the lung diseases such as bronchitis, pneumonia and normal lung images using CT images are presented. The dataset of the lung CT are collected from the radiology department at "Sri Manakula vinayagar medical college and hospital", Madagadipet, Puducherry. The input image is the lung CT obtained from the Philips MX16^{EVO} which uses exclusive technologies such as the EVO Eye algorithm and MAR technology make it possible for the MX16^{EVO} for excellent image quality with reduced noise and deliver performance and productivity that are at the top of its class. The evaluation of the proposed automated diagnosis system of lung diseases are performed by using a set of 300 Lung CT images which is a combination of bronchitis, pneumonia and normal lung images. The original image is converted to gray scale image. Preprocessing technique is applied to enhance the image quality, where three filtering techniques are applied namely the ordered, median and wiener filter. The results obtained from the PSNR values shows median filter results has more PSNR value than the other methods therefore the median filter is applied for preprocessing. The feature extraction is done by the proposed multiscale filter based technique which extracted the filtered co-efficient values. The output of the extracted features are given as an input for feature selection using the evolutionary algorithm technique and classifiers namely MLP-NN, KNN, Bayes net classifiers are used for the classification of the images in to three groups namely bronchitis, pneumonia and normal lung images. The block diagram of the proposed work is represented in Fig.1.

The remainder of this paper is organized as follows. Section 2 describes the related work. Section 3 explains the preprocessing of images. Section 4 deals with the proposed feature extraction technique. Section 5 gives the details of the feature selection using evolutionary algorithm. Section 6 describes the classification of lung diseases using various classifiers. Section 7 explains the results and discussion. Section 8 gives the conclusion. Fig.1 gives the block diagram for the proposed work.

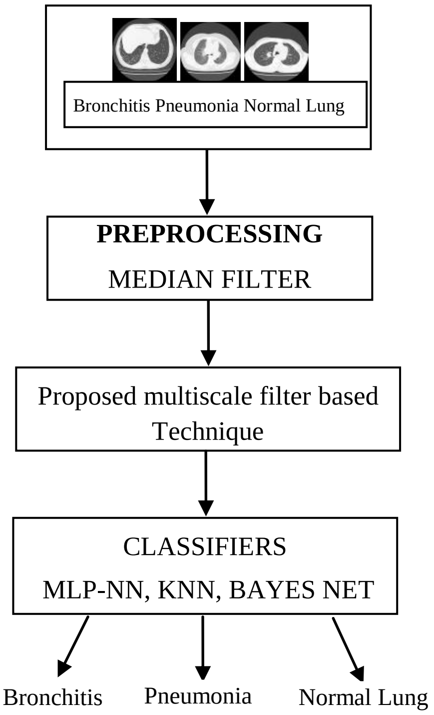

Fig.1. Block diagram of the proposed work

2. RELATED WORK

During the recent period, there are many studies on diagnosis of lung diseases using several features extraction and selection techniques. The model-based methods are based on learning the lung anatomy and developing a model of the lung. Based on this idea, Zrimec and Busayarat [1] first extracted the anatomical features and landmarks. Many modalities of diagnostic imaging (conventional X-ray, computed tomography, nuclear medicine, magnetic resonance imaging, ultrasound scans, etc.) are especially useful because they help to provide effective diagnoses in a noninvasive way. Besides, during the past several years there has been an important increase in the use of diagnostic medical imaging [2]. Computer-aided diagnosis sys-tem is very helpful for radiologist in detection and diagnosing abnormalities earlier and faster [3]. The computer aided diagnosis is a second opinion for radiologists before suggesting a biopsy test [4]. In recent research literature, it is observed that principles of neural networks have been widely used for the detection of lung cancer in medical images [5]. The methodology developed by Xiaomin et al. [6] is divided into three stages. In the first stage, a 2D multi-scale filter is used. In stage 2, blob-shaped nodules and non-nodules are differentiated. In stage 3, the shape features of each region are extracted, and a classifier based on automated rules to reduce false positives is applied. The methodology developed by Sivakumar and Chandrasekar [7] aims to develop a lung nodule detection system by means of segmentation, using fuzzy clustering models and classification with SVM. This methodology uses three types of kernels (linear, poly-nomial and radial basis function (RBF)) for SVM. In recent years, artificial neural networks (ANNs), fuzzy logic and genetic algorithms have been proposed as auxiliary tools in medicine. Fuzzy logic could increase the sensitivity and specificity of biomarkers for lung cancer [8, 9]), whereas ANNs may play an important role in lung cancer, in morphologically differentiating malignant from benign cells [10] and in detection of pulmonary nodules from computed tomography chest images [11]. Lung nodules detection by ensemble classification suggested by Kouzani, et al., [12] achieved lung nodule detection through classification of nodule / non-nodule patterns. It was based on random forests which are ensemble learners growing classification trees. Each tree produces a classification decision. An integrated output was calculated.

3. PREPROCESSING

Preprocessing is needed to improve the quality of the images which to remove noise from the images. Gray scale image conversion is the first step in preprocessing. It makes the feature extraction phase more trustworthy.

After the conversion of the grayscale the filter is applied. The filters that are applicable for the grayscale images are median filter, ordered filter and wiener filter. The output of the various filters is represented as follows. The output of the various filters Peak signals noise ratio (PSNR) represents the median filter has the higher value than other filter. By this justification and the literature the median filter is found to be more appropriate method for the preprocessing technique to remove the salt and the pepper noises from the CT images.

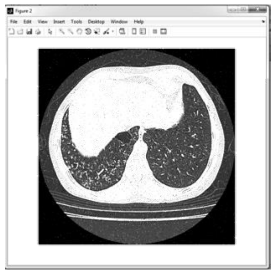

Fig.2. Gray scale image Fig.3. Salt and Pepper noise

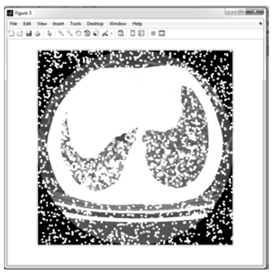

Fig.4. Ordered filter Fig.5. Wiener filter

Fig.6. Median filter

Table.1. PSNR value of filter

SI. No	Filter	PSNR value
1	Ordered filter	20.9974
2	Median filter	28.1666
3	Wiener filter	23.2867

Median filter eliminates the outcome of input noise standards with exceptionally large magnitudes is its major advantage. The median filter formula is

$$y(t) = median(x(t\text{-}T/2), x(t\text{-}T1+1),\ldots,x(t),\ldots,x(t+T/2))$$

where, t represents size of window.

4. FEATURE EXTRACTION

Feature extraction recognizes and extracts remarkable features for a challenging task in order to decrease the complication of processing. It refers to making a division of new features by grouping the existing features. The purpose of feature extraction is to reduce original data set by measuring certain features that distinguish one region of interest from another.

4.1 PROPOSED MULTISCALE FILTER BASED FEATURE EXTRACTION

Modifying the pixels in an image based on some function of a local neighborhood of the pixels refers to image filtering. Gaussian removes high frequency components from the image to low frequency components. Larger σ removes more details. It involves various properties such as commutative, associative, linear, shift variance and differtiation. It is also possible to use second order derivatives to extract the features. A very popular second order operator is the Laplacian operator. The laplacian of a function $f(x, y)$, denoted by, $\nabla^2 f(x, y)$ is defined by:

$$\nabla^2 f(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2}. \qquad (1)$$

The discrete difference approximations are used to estimate the derivatives and represent the Laplacian operator with the convolution mask.

Multiscale filters are just filter banks created using Gaussian functions (whenever e power minus something is Gaussian). Where the goal is to evaluate the capability of different filter banks, based on Gaussian functions. They help to represent image features by giving the response of the image to a particular filter bank. In the proposed work the second order Gaussian derivates are performed followed by the laplacian derivates.

The formula for the second order Gaussian derivatives is given by,

$$G^2(x,y) = \frac{y^2 - \sigma^2 y}{2\pi\sigma_x\sigma_y^5}\exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \qquad (2)$$

and the laplacian transform is given by,

$$LG(x,y) = \frac{\left(x^2 + y^2 - 2\sigma^2\right)}{2\pi\sigma^6}.\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \qquad (3)$$

The proposed method is a combination of Gaussian second order derivatives and laplacian transform. The second order Gaussian derivatives (with aspect-ratio equals 0.25) are oriented at 0, 45, 90 and 135. All the filter banks contain 16 scales .The standard deviation used for the Gaussian-based filter banks is equal to a quarter of the filter-mask size.

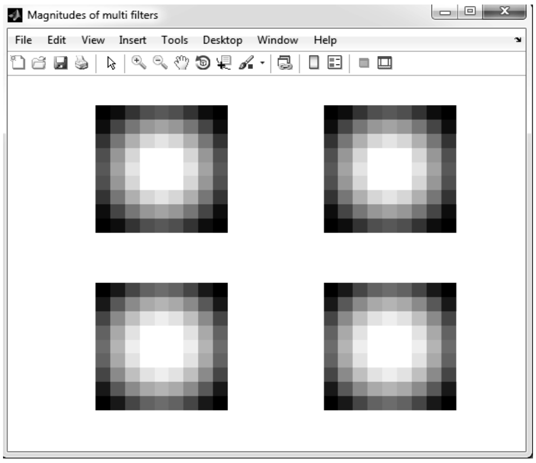

Fig.7. Magnitude of multiscale filter

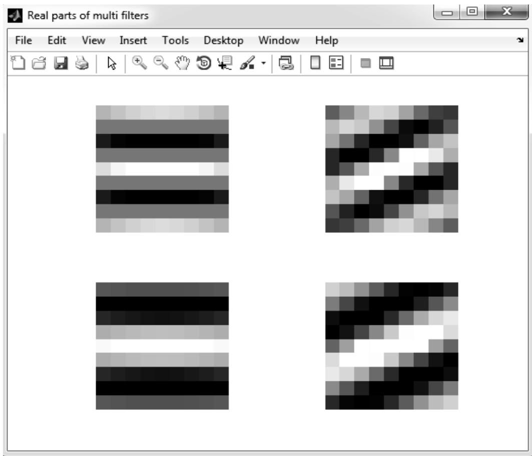

Fig.8. Real part of multiscale filter

The output of the proposed work is the 234 filtered co-efficient values which are extracted from the proposed work. Since it involves both the methods the advantages of the both methods are incorporated in it resulting in new features.

5. FEATURE SELECTION

Feature Selection is a process of choosing minimized applicable features which improves the classification accuracy. Wrapper methods use the classifier as a black box process depending on their predictive power. It removes insignificant features from subsets of features by backward feature elimination. It eliminates the bottom ranked feature from the results.

5.1 GENETIC ALGORITHM

Genetic Algorithm is a population based search methods and it moves from one set of points (population) to another set of points in a single iteration with likely improvement using set of control operators. It improved fitness in the course of evolution. Chromosome consists of a set of elements, called 'genes', which holds a set of values for the optimization variables that provides the solution to a given problem. For the initial population the individuals are assessed through a fitness function, for optimizing task which can be minimization or maximization in a genetic algorithm. Based on this fitness function, chromosomes are selected and genetic operators like mutation, crossover is applied to the selected chromosomes. Offspring for the next generation are formed by using two main genetic operators, crossover and mutation. Crossover combines the features of two individuals to create two similar offspring by randomly selecting a point in the two selected parents gene structures and exchanging the remaining segments. Mutation operates by randomly changing one or more components of a selected individual. This operator prevents any stagnation that might occur during the search process. Genetic algorithms have established extensive improvement over a variety of random and local search methods. GA initially starts with unknown search space in order to bias subsequent search into promising subspaces. These chromosomes evolve, better individuals till it reaches the global optimum solution.

The proposed genetic algorithm states the initial population is assigned as 30 which is combined via a crossover operator to produce offspring. Two point crossover with crossover rate of 0.7. The individuals in the population are then evaluated via a fitness function and the process of crossover, evaluation, and selection is repeated for a predetermined number of generations or until a satisfactory solution has been found or condition fails. In the feature selection formulation of the genetic algorithm, individuals are composed of bit strings: a 1 in bit position indicates that feature should be selected 0 indicates this feature should not be selected. The mutation rate is 0.01 and the stopping condition is RMSE threshold of 0.001 or 500 operations with generation of 500. Fitness condition for RMSE and the crossover operator is two points. The top ranking 30 features are selected and fed into the classifiers.

6. CLASSIFIERS

A classifier assigns one class to each point of the input space. The input space is thus partitioned into disjoint subsets, called decision regions, each associated with a class.

6.1 MULTILAYER PERCEPTRON – NEURAL NETWORKS

A multilayer perceptron (MLP) is a feed forward neural network with input layer of source neurons, one or more hidden layers or computational neurons, and an output layer of computational neurons. The input signals are spread in a forward direction on a layer-by-layer basis. A hidden layer "hides" its desired output. Each layer can contain from 10 to 1000 neurons. Neurons in the hidden layer cannot be practical through the input/output behaviour of the network. A training set of input patterns is presented to the network. The network computes its output pattern, if it is an error free – or difference between actual and desired output patterns – the weights are adjusted to reduce this error. In a back-propagation neural network, the learning algorithm has two stages. First, a training input pattern is given to the network input layer. The network propagates the input pattern from layer to layer until the output pattern is generated by the output layer. If the pattern is different error is propagated from the output layer to the input layer. The weights are modified as the error is propagated.

1) Initialisation: Initialise the weights and threshold levels of the association to random numbers uniformly
2) Activation: Activate the back-propagation neural network by applying inputs $x_1(p)$, $x_2(p)\ldots$, $x_n(p)$ and desired outputs $y_{d,1}(p)$, $y_{d,2}(p),\ldots$, $y_{d,n}(p)$.

$$y_j(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p).w_{ij}(p) - \theta_j\right] \quad (4)$$

where, n is the number of inputs of neuron j in the hidden layer, and *sigmoid* is the sigmoid activation function. Calculate the actual outputs of the neurons in the hidden layer:

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{jk}(p).w_{jk}(p) - \theta_k\right] \quad (5)$$

where, m is the number of inputs of neuron k in the output layer.

3) Weight training:
    a) Update the weights in the back-propagation network propagating backward the errors associated with output neurons, calculate the weight corrections, update the weights at the output neurons.
    b) Calculate the error gradient, weight corrections and update the weights at the hidden neurons.
4) Iteration:

    Increase iteration p by one, go back to step 2 and repeat the process until the selected error criterion is satisfied.

In this work the number of features is equal to number of input neurons, the number of neurons in the hidden layer is one third of the number of neurons in the input layer, the number of classes constituted the number of neurons in the output layer, the Sigmoidal activation function with the initial random weights between 0 and 1 were used and the bias value was set to zero finally the output will result in 3 classes of labels namely bronchitis, pneumonia and normal lung.

6.2 K NEAREST NEIGHBOUR

Nearest Neighbor was proposed by Fix and Hodges (1951). It has a set of stored records, Distance Metric to compute distance between records, the value of k, the number of nearest neighbors is to be retrieved. Distance (k) is computed to classify an unknown record and use class labels of nearest neighbors to determine the class label of unknown record. It is done by Euclidean space, classification is belated till a new instance arrives, classification is done by comparing feature vectors of the different points, final function may be discrete or real-valued. The Advantage is Simple, Powerful, Requires no training time and nonparametric architecture.

The KNN classifier works as follows:

1) Compute distance between two points: In cartesian coordinates, if $a = (a_1, a_2,..., a_n)$ and $b = (b_1, b_2,..., b_n)$ are two points in Euclidean n-space, then the distance from a to b, or from a to b is given by Euclidean distance,

$$d(a,b)=\sqrt{\sum_i (a_i - b_i)^2} \tag{6}$$

2) Find the class from nearest neighbor list
   - The k-nearest neighbors
   - Weigh the vote according to distance where weight factor, $w = 1/d^2$
3) Choosing the k value of:
   - If k is too small, sensitive to noise points
   - If k is too large, neighborhood may include points from other classes
4) Kd-tree construction algorithm
   - Select the x or y dimension
   - Partition the space into two
   - Repeat recursively in the two partitions as long as there are enough points

The 100 attributes and class index of 125 are taken. The values of the column are taken as 1 to 16, 17 to 32, 33 to 48, 49 to 64, 65 to 80 ,81 to 100 order and the values of the class label are computed.

6.3 BAYES NET

A Bayesian network, $B =< N, A, \odot >$ is a directed acyclic graph (DAG) $<N, A>$ with a conditional probability distribution for each node, collectively represented by $\odot$. Each node $n \in N$ represents a domain variable, and each arc $a \in A$ between nodes represents a probabilistic dependency. The pair (G, CPD) encodes the joint distribution. It is a combination of the bayes network and decision tree based method. In general, a BN can be used to compute the conditional probability of one node, given values assigned to the other node, hence, a BN can be used as a classifier that gives the posterior probability distribution.

7. RESULTS AND DISCUSSIONS

The input CT images are preprocessed and the features are extracted and the selection of the features are done by the evolutionary approach and fed in to the classifiers. The classification accuracy is found for each classifier. The testing and the training of the dataset are done in tenfold cross validation.

True positive refers to correctly identified instances in a dataset. False positive refers to incorrectly identified instances in a dataset. True negative refers to correctly rejected instances in a dataset. False negative refers to incorrectly rejected instances in a dataset.

Classification Accuracy refers to the number of accurately classified instances. The classifiers such as MLP NN, KNN and Bayes net and the classification accuracy are represented in the Fig.9.

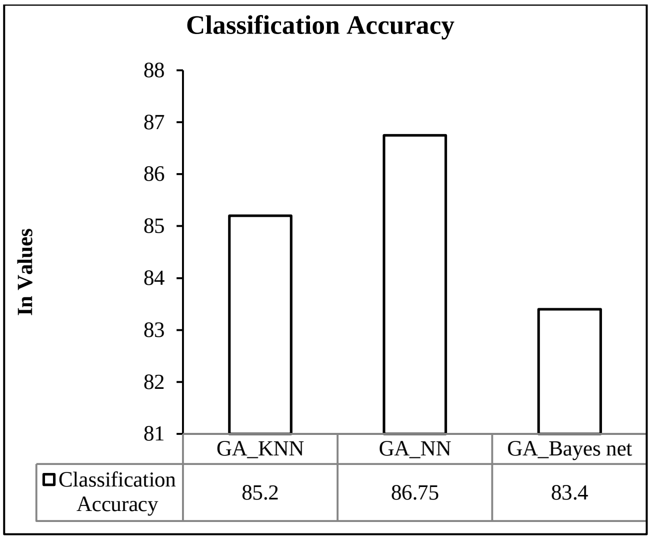

Fig.9. Classification Accuracy

Sensitivity measures the proportion of actual positives which are identified as such. Specificity measures the proportion of negatives which are correctly identified as such. F-measure - is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. Positive predictive value is the probability that subjects with a positive screening test.

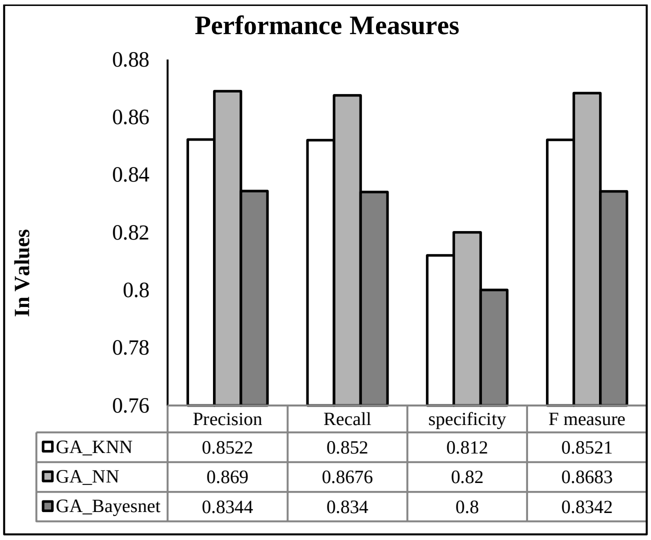

Fig.10. Performance Measures

The dataset of the three diseases are taken and the classification of the diseases are represented as

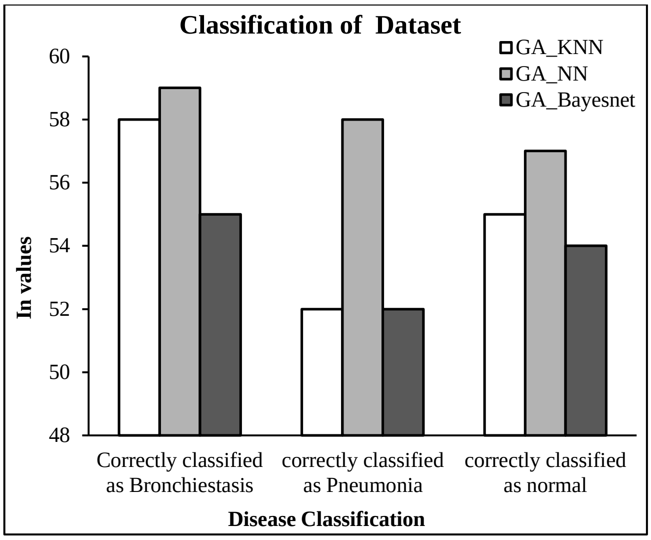

Fig.11. Classification of Dataset

The results and the discussion shows the various performance measures and the results obtained for the dataset.

8. CONCLUSION

Lung diseases are one of the major health issues worldwide. The research work "A novel shape based feature extraction technique for diagnosis of lung diseases using evolutionary approach", inputs the lung CT images of three dataset namely bronchitis, pneumonia and normal lung. The preprocessing to enhance the image are done by median filter and the feature are extracted by proposed multiscale filter based technique and the feature selection is evolved by the genetic algorithm and the classifiers such as MLP-NN, KNN and Bayes net are used to classify the dataset in to 3 classes, the results ensures that the MLP-NN, KNN, Bayes net has an accuracy of 86.75 %, 85.2 %, 83.4% of classification accuracy. Thus MLP-NN classifier acts as a best classifier for classifying the lung diseases in this work. The enhancement of the proposed work can be done by adding more features extraction, selection techniques and different classifiers can also be included. Hence this work assists the doctors in the medical field for diagnosing the lung diseases using soft computing and evolutionary computing techniques.

ACKNOWLEDGEMENT

We would like to acknowledge Radiologist Dr. Ramesh Kumar M.D., (R.D), Professor and Head, Department of Radiology, Sri Manakula Vinayagar Medical College and Hospital Madagadipet for providing the CT images.

REFERENCES

[1] T. Zrimec and S. Busayarat, "A system for computer aided detection of disease patterns in high resolution CT images of the lungs", *IEEE International Symposium on Computer-Based Medical Systems*, pp. 41-46, 2007.

[2] R. Smith-Bindman, D. L. Miglioretti and E. B. Larson, "Rising use of diagnostic medical imaging in a large integrated health system", *Health Affairs*, Vol. 27, No. 6, pp. 1491-1502, 2008.

[3] R. N. Strickland, "Tumor detection in non stationary backgrounds", *IEEE Transactions on Medical Imaging*, Vol. 13, No. 3, pp. 491-499, 1994.

[4] S.-C.B. Lo, S.-L.A. Lou, Jyh-Shyan Lin, M. T. Freedman, M. V. Chien and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection", *IEEE Transactions on Medical Imaging*, Vol. 14, No. 4, pp. 711-718, 1995.

[5] G. Coppini, S. Diciotti, M. Falchini, N. Villari and G. Valli, "Neural networks for computer aided diagnosis: detection of lung nodules in chest radiograms", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 7, No. 4, pp. 344-357, 2003.

[6] Pei Xiaomin, Guo Hongyu and Dai Jianping, "Computerized detection of lung nodules in CT images by use of multiscale filters and geometrical constraint region growing", *Fourth International Conference on Bioinformatics and Biomedical Engineering*, pp. 1-4, 2010.

[7] S. Sivakumar and C. Chandrasekar, "Lung nodule detection using fuzzy clustering and support vector machines", *International Journal of Engineering and Technology*, Vol. 5, No. 11, pp. 179-185, 2013.

[8] J. Schneider, N. Bitterlich, H. G. Velcovsky, H. Morr, N. Katz and E. Eigenbrodt, "Fuzzy logic-based tumor-marker profiles improved sensitivity in the diagnosis of lung cancer", *International Journal of Clinical Oncology*, Vol. 7, No. 3, pp. 145-151, 2002.

[9] J. Schneider, G. Peltri, N. Bitterlich, H. G. Velcovsky, H. Morr, N. Katz and E. Eigenbrodt, "Fuzzy logic-based tumor marker profiles improved sensitivity of the detection of progression in small-cell lung cancer patients", *International Journal of Clinical Oncology*, Vol. 2, No. 4, pp. 185-191, 2003.

[10] Zhi-Hua Zhou, Yuan Jiang, Yu-Bin Yang and Shi-Fu Chen, "Lung cancer cell identification based on artificial neural network ensembles", *Artificial Intelligence in Medicine*, Vol. 24, No. 1, pp. 25-36, 2002.

[11] M. Gomathi, and P. Thangaraj, "A computer aided diagnosis system for lung cancer detection using machine learning technique", *European Journal of Scientific Research*, Vol. 51, No. 2, pp. 5770-5779, 2011.

[12] A. Z. Kouzani, S. L. A. Lee and E. J. Hu, "Lung nodules detection by ensemble classification", *IEEE International Conference on Systems, Man and Cybernetics*, pp. 324-329, 2008.

[13] Rajdev Tiwari and Manu Pratap Singh, "Correlation-based Attribute Selection using Genetic Algorithm", *International Journal of Computer Applications*, Vol. 4, No. 8, pp. 28-34, 2010.

[14] I. H. Witten and E. Frank, "*Data Mining: Practical machine learning tools and techniques*", Second Edition, Morgan Kaufman, 2005.

[15] H. D. Tagare, C. C. Jafe and J. Duncan, "Medical image databases: A content-based retrieval approach", *Journal of the American Medical Informatics Association*, Vol. 4, No. 3, pp. 184-198, 1997.

[16] Nassir Salman, "Image Segmentation Based on Watershed and Edge Detection Techniques", *The International Arab*

Journal of Information Technology, Vol. 3, No. 2, pp.104-110, 2006.

[17] A. Jain and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 2, pp. 153-158, 1997.

[18] M. J. Martin-Bautista and M-A Vila, "A Survey of Genetic Feature Selection in Mining Issues", *Proceedings of the Congress on Evolutionary Computation*, Vol. 2, pp. 1314-1321, 1999.

[19] Rafael C. Gonzalez, Richard E. Woods, "*Digital Image Processing*", Third Edition, Pearson Prentice Hall, 2008.

[20] S. N. Sivanandan, S. Sumathi and S. N. Deepa, "*Introduction to Neural Networks Using Matlab*", Tata McGraw Hill Publishing Company Limited, 2006.