International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-4 Number-1 Issue-14 March-2014

Classification of Lung Diseases by Image Processing Techniques Using Computed Tomography Images

C.Bhuvaneswari[1], P.Aruna[2], D.Loganathan[3]

Abstract

Lung diseases are the disorders that affect the lungs, the organs that allow us to breathe and it is the most common medical conditions worldwide especially in India. The diseases such as pleural effusion and normal lung are detected and classified in this work. The purpose of the work is to detect and classify the lung diseases by effective feature extraction through moment invariants, feature selection through genetic algorithm and the results are classified by the Naïve bayes and decision tree classifiers. The preprocessing techniques will remove the noises and the feature extraction are done to extract the useful features in the image and the feature selection technique will optimize the top ranking features that are relevant for the image and the classifiers are employed to classify the images and the performance measures are found for the same. The result shows that the Decision tree classifier shows more promising results than the naïve bayes classifier.

Keywords

Classification, Decision tree, Feature Extraction, moment invariants, Preprocessing, Performance.

1. Introduction

The term lung disease refers to many disorders affecting the lungs such as asthma, chronic obstructive pulmonary (COPD) disease, infections such as tuberculosis, influenza, lung cancer, pneumonia and other breathing problems. Lung diseases signs and symptoms can differ by the type of the affected disease. Common signs are trouble in breathing, shortness of breath, feeling like you're not getting enough air, decreased ability to exercise, a cough that won't go away, coughing up blood or mucus, pain or discomfort when breathing in or out. Medical image analysis and process requires an environment for data access, data analysis, processing, revelation and algorithm development. Medical imaging is the technique and process used to create images of the human body for clinical purposes for diagnosis and analysis or medical science (including the study of disease of normal anatomy and physiology).In this paper, an automated approach for classification of the lung diseases using CT images is presented. The lung CT image is engaged as the input. Another look of the proposed system of lung diseases detection has been performed by using a set of CT images (pleural effusion and normal lung). The original image is transformed to gray scale image. After that, removal of the noises and contrast enhancement is done for obtaining the enhanced images. The median filter is applied to remove the salt and the pepper noises and the preprocessed images are given as input for feature extraction where the useful features of the images are extracted and the extracted features are selected by the genetic algorithm method, the classifiers are used to classify the datasets in to relevant datasets and the performance measures are evaluated for the datasets.
The paper is organized as Section 2 deal with the literature review. Section 3 explains the projected work. In the section 4 experiments and results is detail and section 5 deals with the results of the findings. This works helps as the assistant in detecting the lung diseases in the medical field .the motivation for choosing the work is to give the automated way of finding the diseases that affect our community most commonly.

2. Literature Review

In 1962 Hu proposed the theory of algebraic invariants probably originates from famous German mathematician David Hilbert [1] and was thoroughly studied also in [2], [3]. Moment invariants were firstly introduced to the pattern recognition community [4], who employed the outcome of the algebraic invariants and consequent his seven renowned invariants to rotation of 2-D objects.

Manuscript received on January 25 , 2014.
C.Bhuvaneswari, Head, Department of computer science, Theivanai ammal college for Women(Autonomous),Villupuram, India.
P.Aruna, Professor, Department of computer science and Engineering, Annamalai University, Chidambaram, India.
D.Loganathan, Professor and Head, Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India.

Image processing techniques[5][6][7]can be regarded as Low level processing where we gives input is a image and the output is also a image and includes noise removal and image sharpening. Middle level image processing where image is an input but output will be aspect processes like object recognition, segmentation. Uppaluri et al. [8] have developed a general system for regional classification by using small areas that are classified into six categories based on 15 statistical and fractal texture features.

Manish Kakar et al., [9] proposed a method based upon the, Gabor filtering to extract texture features, From the segmentation results, the accuracy of delineation was seen to be above 90%.For recognizing the segmented regions automatically, an average sensitivity of 89.48%was obtained by combining, shape position-based features and cortex-like feature in a Simple SVM classifier. Summers proposed that is a time-costly process, and the quantity of images to be examined is at an unmanageable level, especially in populous countries with scarce medical professionals. While examine the X ray image by the radiologists, they need to recognise the two lungs and then find any obvious abnormality. CAD systems are usually specialized in anatomical regions such as the thorax, breast, or colon by using certain medical imaging technologies such as radiography, CT or magnetic resonance imaging (MRI) [10]. Kim Ko, and Jung (1993) proposed a new method based on neural networks to solve the DFR problem, including time intervals. Computed Tomography (CT) is efficient than X-ray [11,12,13,15], the latter is more generally available. Thus initial diagnosis for TB and lung cancer, now performed by medical doctors, is mainly based on chest X-ray images.

3. Proposed work

For the classification of lung diseases by image processing techniques using computed tomography images the following proposed model is evolved for obtaining the desired results. The proposed work is developed by Matlab 10.

The model for the proposed work is as follows

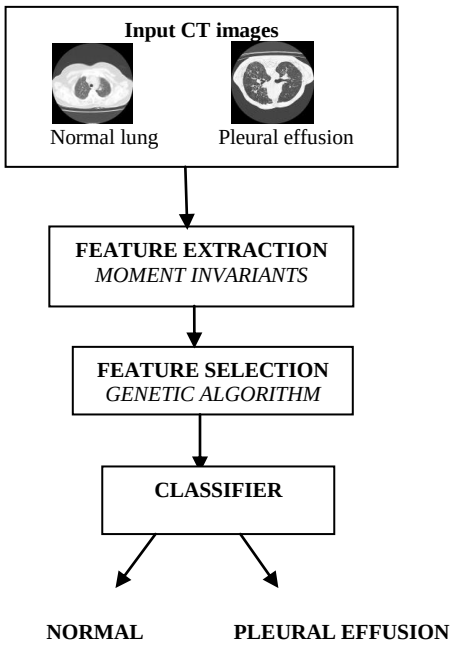


Fig 1: Block diagram of the proposed work

The following steps are involved in the proposed work

Step1: In order to prepare the image for classification the input images are preprocessed by the median filter to remove the salt and the pepper noises.

Step 2: The feature extraction process is done to extract the relevant features from the images. The seven moment invariants are derived from the theory of algebraic invariant.

Step 3: The Genetic Algorithm is applied for the feature selection technique for finding the optimal top ranking features.

Step 4: The Classification of the images are done by the Naïve Bayes and J48 decision tree Classifiers to classify the to classify the two class problem as normal lung image or Pleural effusion.

Step 5: Performance measure of Precision Recall-measures are computed.

3.1 Preprocessing

Image pre-processing can significantly increase the reliability of an optical inspection.

a. The CT lung disease image is taken as an input image.

b. The image is resized to 128 by 128 of 16 x16 windows and grey scale image is extracted.

c. Median filter technique is applied to remove the noise from the images.

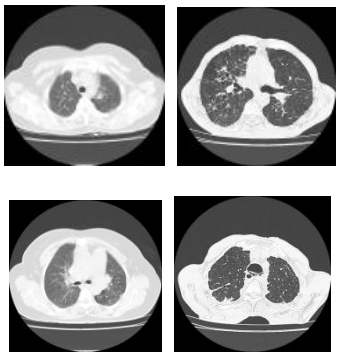
It is a nonlinear operation used in image processing to reduce "salt and pepper" noise. A median filter is useful than convolution goal is to concurrently reduce noise and preserve edges. An image with salt-and-pepper noise will have bright pixels in dark regions and dark pixels in bright regions. These types of noise are occurred by dead pixels, analog-to-digital converter errors, bit errors in transmission, etc.

The median filter is given by

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

Where t is the size of the window of the median filter.

This can be removed by large part by using dark frame subtraction and by interpolating around bright/dark pixels.



(a) (b)

Fig 2: Input image (a) normal lung (b) pleural effusion

3.2 Feature Extraction

Feature extraction is performed to reduce the input vector size. The input data is to be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. It refers to ttransforming the existing features into a lower dimensional space. The purpose of feature extraction is to reduce original data set by measuring certain features that distinguish one region of interest from another.

3.2.1 Moment invariants

Moment invariants are a set of nonlinear functions, which are invariant to translation, scale, and orientation and are defined on normalized geometrical central moments. Hu first introduced seven moment invariants based on normalized geometrical central moments up to the third order. Moments gives description of an object that exceptionally represents its shape. Invariant shape recognition is done by classification in the multidimensional moment invariant feature space.Hu defines seven of the shape descriptor values that are computed from central moments through order three that are independent to object translation, scale and orientation. Translation invariance is obtained by computing moments that are normalised with respect to the centre of gravity resulting in the distribution at the origin from the centre of mass (central moments). Size invariant moments are imitative from algebraic invariants but these can be exposed to be the result of a simple size normalisation. The second and third order values obtained from the normalised central moments a set of seven invariant moments can be computed which are independent of rotation.

The formula for the calculation of the moment invariants

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \mu_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \mu_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \quad \textbf{(2)}$$

The seven moment invariants are useful properties of being unchanged under image scaling, translation and rotation.

Steps involved in moment invariants feature extraction Techniques:

Step1: Calculate the moment vector of the image.
Step 2: Calculate the central moments of the image.
Step 3: Calculate the centroid of the image
Step 4: Calculate the geometric moment of the image.
Step 5: Calculate the seven moment invariants.

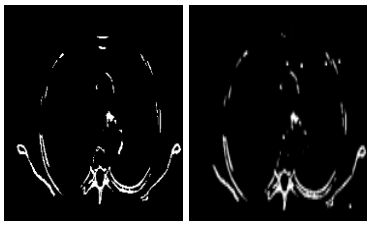


Fig 3: Moment Invariant images

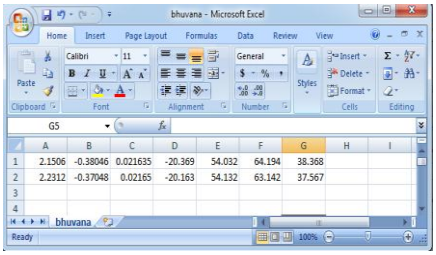


Fig 4: Feature extracted values in .csv file

3.3 Feature selection

The term Feature selection [14] works with selecting a subset of features, among the entire features, that shows the best performance in classification accuracy. Optimisation process is done by the genetic algorithm which is the efficient methods for function minimization. A genetic algorithm mainly composed of three operators: selection, crossover, and mutation. In selection, a good string is selected to breed a new generation, crossover combines good strings to generate better offspring. The mutation changes a string locally to maintain genetic diversity from one generation of a population of chromosomes to the next. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not fulfilled, the population is operated upon by the three GA operators and then re-evaluated.

The pseudo code for the genetic algorithm is

```
BEGIN
  Generate initial population;
  Compute fitness of each individual;
  REPEAT /* New generation /*
    FOR population size DO
      Select two parents from old generation;
       /* biased to the fitter ones */
      Recombine parents for two offspring;
      Compute fitness of offspring;
      Insert offspring in new generation
    END FOR
  UNTIL population has converged
END
```

The genetic algorithm is implemented in the matlab with the entropia function will compute the values essential for the ranking. The statistics functions as entropy of each will compute the output entropy followed by mutual information between features and output finally the mutual information between features are calculated.

The feature selector will select the parameters, with the starting genes then loop over the generations rank the features, find the relevance, remove redundancy and fix the fitness function then rearrange the population according to their fitness and also find the crossover then verify the not used features and repeated features and features without entropy. After the application of the genetic algorithm the most relevant 30 features are ranked optimally and fed in to the classifiers.

3.4 Classifiers

To distinguish between different types of objects, the two classifiers are used to classify the images and the performance measures of them are calculated.

3.4.1 Naive Bayes classifier

The Naive Bayes classifier is used for classifying the images and the performance measures are calculated and the comparative study of the successful feature selection algorithm is chosen. Naive Bayes is used for classifying the selected features in this work. The selected features are classified to the most likely class. Learning in Naïve Bayes is simplified by considering the features that are independent for a given class. The feature is classified as shown in equation

$$P(X|C) = \Pi_{i=1}^{n} P(X_i|C) \quad (3)$$

Where X=(X1,…, Xn) refers the feature vector and C is a class. The selection of choosing this classifier is the feature selection method used in this study prefers a suitable classifier that handles all type of value. The experiments and results show the outcome of the work done by the classifier and the results are shown.

3.4.2 Decision Tree

A decision tree is a predictive machine-learning model that decides the target value of a new sample based on a variety of attribute values of the existing data. The internal nodes of decision tree represent the diverse attributes, the branches linking the nodes gives the possible value, the terminal nodes tell us the final value (classification) of the dependent variable. The dependent variable are the attributes which are predicted, as its value depends upon, or is decided by, the values of all the other attributes. Independent variables are the values that predict the other variables.The J48 Decision tree classifier has the following steps. In order to categorize a new item, it first needs to create a decision tree based on the attribute values of the existing training data. If the training set is identified it finds the attribute that differentiate the instances in a systematic manner. This feature will ensure the data instances to identify

and classify as the best which has the highest information gain. The branches are terminated if among the possible feature values without ambiguity if the data instances lie inside the category and have the same value for target variable and assign to it the target value that we have obtained.

4. Experiment and Results

In order to prepare the image pre-processing of the image was done by contrast enhancement and median filtering. Median filter was done for noise removal. Contrast enhancement was performed to get the clear image. The preprocessed images are given as a input to the feature extraction process, the statistical features are extracted from the preprocessed images. The feature selection is done by the genetic algorithm where the relevant features are extracted and the images are classified using the classifiers. The evaluation of the testing instances using the moment invariants with naive bayes and decision tree is tabulated below.

The following steps are followed in the proposed work

Step 1: The input CT image of 128x128 pixels are given as input.

Step 2 : The preprocessing was done by the median filter and the salt and the pepper noises are removed to get the enhanced image.

Step 3: The feature extraction of the moment invariants are calculated which results in seven geometric features extracted from each image.

Table 1: Evaluating the testing instances

S.No	Performance measure	NaïveBayes Class I	NaïveBayes Class II
1	TP Rate	1	0.75
2	FP Rate	0.25	0
3	Precision	0.75	1
4	Recall	1	0.75
5	F-Measure	0.86	0.86
6	ROC Area	1	1

Step 4: The feature selection is done using the evolutionary approach of the genetic algorithm where the top ranked 30 features are evaluated and given as an input to the classifiers.

Step 5: The naïve Bays and the decision classifiers are applied to calculate the performance measures of the two class problem.

The performance measures such as The *True Positive (TP)* rate is the proportion of datasets which were classified as class *a*, among all datasets which truly have class a, i.e. how much part of the class was achieved which is equal to *Recall*.

The *False Positive (FP)* rate is the proportion of examples which were classified as class a, but belong to a diverse class, among all examples which are not of class *x*.

Table 2: Detailed Performance Accuracy for classes using Naive bayes

S. No	Evaluation of testing instances	Naïve Bayes	Decision tree
1	Correctly Classified Instances	6	7
2	Incorrectly Classified Instances	1	0
3	Kappa statistic	0.72	1
4	Mean absolute error	0.1429	0.0048
5	Root mean squared error	0.378	0.0073
6	Relative absolute error	29.13 %	0.97%
7	Root relative squared error	76.37 %	1.4698%
8	Coverage of cases (0.95 level)	85.71 %	100%
9	Mean relative region size (0.95 level)	50%	50%

The *Precision* is the part of the examples which truly have class *x* among all those which were classified as class *x*.

Precision (P) = tp/(tp+fp)

the recall is defined by **Recall (R) = tp/(tp+fn)** Where tp, fp, tn and fn are true positive, false positive, true negative and false negative respectively. The *F-Measure* is simply 2*Precision*Recall/(Precision+Recall) are calculated and displayed The experimental results of the above proposed work in the work and the comparative results are tabulated.

Table 3: Detailed Performance Accuracy for classes using Decision tree

S.No	Performance measure	Decision tree Class I	Decision tree Class II
1	TP Rate	1	1
2	FP Rate	0	0
3	Precision	1	1
4	Recall	1	1

5	F-Measure	1	1
6	ROC Area	1	1

The detailed performance measure of the classifiers and their results are tabulated and the results shows for this set of the dataset the decision tree classifier gives more accurate results than the naïve Bayes classifier.

5. Conclusion and Future Work

In this work the pre-processing of the images are done then the feature extraction is done by moment invariants and the feature selection is done by the genetic algorithm and the naïve bayes and the decision tree classifiers are used to train, test and classify and the performance measure shows that the decision tree classifier produces more relevant results than the naïve bakes classifier. This work may be enhanced by taking some more lung diseases and more feature extraction methods can be employed and by combining with other classification models for obtaining more accurate results.

Acknowledgment

We acknowledge Dr. Ramesh Kumar M.D.,(R.D) ,Professor and Head, Department of Radiology Sri Manakula Vinayagar Medical college and Hospital Madagadipet for analyzing the Dicom Images regarding the lung diseases.

References

[1] D. Hilbert, "Theory of Algebraic Invariants", Cambridge University Press, 1993.
[2] G. B. Gurevich, "Foundations of the Theory of Algebraic Invariants", Groningen, The Netherlands: Nordhoff, 1964.
[3] Cui, F.-y.; Zou, L.-j; Bei Song, IEEE International Conference on," Edge feature extraction based on digital image processing techniques**",** Page(s):2320 – 2324, 1-3 Sept. 2008.
[4] M. K. Hu, "Visual pattern recognition by moment invariants,", IRE Transaction Information Theory, vol. 8, pp. 179–187, 1962.
[5] Gonzalez, R.C. and R.E. Woods, "Digital Image Processing" Reading, Massachusetts: Addison-Wesley. 716, 1992.
[6] Abby A. Goodrum" Image information retrieval An overview of current research", special issue on information science research, volume 3, No 2 , 2000.
[7] Giardina, C.R. and E.R. Dougherty, "Morphological Methods in Image and Signal Processing", Englewood Cliffs, New Jersey: Prentice–Hall. 321, 1988.
[8] Uppaluri R, Hoffman EA, Sonka M, Hartley PG, Hunninghake GW, McLennan G,"Computer recognition of regional lung disease patterns" American Journal of Respiratory Critical Care Medicine 1999;160:pg:648–54.
[9] Manish Kakara, Dag Rune Olsen "Automatic segmentation and recognition of lungs and lesion from CT scans of thorax " IEEE transactions on Computerized Medical Imaging and Graphics 33 (2009),pg: 72–82.
[10] R. M. Summers, "Road maps for advancement of radiologic computer-aided detection in the 21st century," Radiology, vol.229, no. 1, pp. 11–13, 2003.
[11] Azadeh Bastani et al. ,"A comparision between walshhadamard and fourier analysis of the eeg signals", International Journal of Engineering Science and Technology (IJEST), ISSN : 0975-5462 Vol. 3, No. 7, July 2011, pg: 5470-5476.
[12] Kakeda, S. et al., "Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer-Aided Diagnosis System", American Journal of Roentgenology, 182, February, pp. 505-510, 2004.
[13] H.D.Tagare, C. Jafe, J. Duncan, "Medical image databases: A content-based retrieval approach", Journal of the American Medical Informatics Asssociation,4 (3),1997, pp. 184-198.
[14] A. Jain and D. Zongker, "Feature Selection: Evaluation,Application, and Small Sample Performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, Vol. 19, No. 2, pp. 153-158.
[15] Lee Y, Hara T, FujitaH, Itoh S, Ishigaki T.,"Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique". IEEE Transaction Medical Imaging 2001; 20:595–604.

C.Bhuvaneswari received her Bachelor degree from Madras University, Post graduate from Bharathidasan University and Masters of Philosophy from Madurai Kamaraj University, Madurai. At present she is pursuing Ph.D at Annamalai University,Chidambaram. She has nine years of teaching experience and 5 years of research experience. Her main research areas include image processing and data mining. She has published five research papers in international journals and six papers in international conferences and ten papers in national conferences.

Dr. P.Aruna received the B.E. from Madras University, M.Tech from IIT Delhi. She received the Ph.D. degree from the Annamalai University. Currently, she is a professor at Annamalai University, Chidambaram. She has published twenty research papers in International journals and five research papers in national journals. She has published fifteen research papers in International conferences and twenty two research papers in national conferences. She has twenty two years of teaching experience and thirteen years of experience. Her research interests include Neural networks & Fuzzy systems, Data Mining and Image processing .

D.Loganathan received the Post-Graduate Degree from Birla Institute of Technology and Science, and obtained Doctorate from Anna University, Chennai. Currently, he is a professor and Head in the Department of Computer Science and Engineering at Pondicherry Engineering College, Puducherry, INDIA. His research interests are information security, image processing. Earlier deputed for TEN week training at CICC Tokyo, Japan. He has been associated with a number of National , International Conferences and workshops and acted as Panel Member for AICTE and NBA. He is also associated with many Professional bodies.