



Analysis of Computed Tomography Imaging Signs of Lung Diseases Based on Modified Genetic Optimization Using Fisher Criterion

Thushara Lalg¹, Sreejakumari²

PG Student [AE], Dept. of ECE, Sivaji College of Engineering and Technology, Manivila, Tamilnadu, India¹

Assistant Professor, Dept. of ECE, Sivaji College of Engineering and Technology, Manivila, Tamilnadu, India²

ABSTRACT: The classification and selection of the optimal features from the common Computed Tomography (CT) imaging signs of lung diseases (CISLs) improves the performance of the classification process and reduces the time complexity and algorithm complexity of the overall process. Computed Tomography (CT) images are used for the identification of the diseases in the lung images. Features are extracted from the lung images based on B-HOG features, Wavelet features, Local binary pattern (LBP) features and CVH features. The feature selection process are employed based on the Genetic algorithm in which the Fisher criterion is employed for the objective function and to employ best fitness function based on the fisher criterion, select the optimal features and combine fitness function with Genetic optimization in order to improve the efficiency of Genetic algorithm. After the mutation and the crossover steps, the stopping condition is verified and the process is repeated till the stopping condition is reached. The selected features are then classified using different classifiers such as Support vector machine (SVM), Bag of Features, Naive Bayes (NB), k-nearest neighbour (k-NN), AdaBoost (Ada), in order to compare the performance of the classifiers. The performance of the process are measured based on the performance metrics like Accuracy, Sensitivity and Specificity.

KEYWORDS: Common CT imaging signs of lung diseases (CISLs), feature selection, fisher criterion, feature extraction, lung lesion classification, medical image classification.

I.INTRODUCTION

In this paper, it proposes texton signatures based on raw pixel representation along with a parallel multiple classifier system for the classification of emphysema in computed tomography images of the lung. The multiple classifier system is composed of support vector machines on the texton signatures as base classifiers and combines their decisions using product rule. The proposed approach is tested on 168 annotated regions of interest consisting of normal tissue, centrilobular emphysema, and paraseptal emphysema. Texton-based approach in texture classification mainly has two parameters, i.e., texton size and k value in k-means. Our results show that while aggregation of single decisions by SVMs over various k values using multiple classifier systems helps to improve the results compared to single SVMs, combining over different texton sizes is not beneficial. The performance of the proposed system, with an accuracy of 95%, is similar to a recently proposed approach based on local binary patterns, which performs almost the best among other approaches in the literature.

In this study, texture based segmentation and recognition of the lung diseases from the computed tomography images are presented. The texture based features are extracted by Gabor filtering, feature selection techniques such as Information Gain, Principal Component Analysis, correlation based feature selection are employed with Genetic algorithm which is used as an optimal initialization of the clusters. The feature outputs are combined by watershed segmentation and the fuzzy C means clustering. The images are recognized with the statistical and the shape based features. The four classes of the dataset of lung diseases are considered and the training and testing are done by the Naive Bayes classifier to classify the datasets. Results of this work show an accuracy of above 90% for the correlation based feature selection method for the four classes of the dataset. Lung diseases are leading cause for the most disabilities and death in the world. Radiologist diagnosis the chest CT and the success of the radiotherapy depend on the dosage of the drugs given and the doses that affect the normal tissues surrounding areas.



The chest CT shows the first important modality of the assessment of the diseases. The CT image along with the symptoms of the diseases will give detailed assessment about the lung diseases. The major causes of the lung diseases are caused by smoking, inhaling the drugs, smoke and allergic materials. The lung diseases are generally identified by the symptoms and the regular dosage of the antibiotics may cure the disease. If the antibiotics does not respond to the disease the computed tomography images assists in detecting the severity of the lung diseases. There are many types of the disease that causes the lung infection such as inflammatory lung diseases, chronic obstructive pulmonary disease(COPD), Emphysema, Chronic Bronchitis, pleural effusion, Intersitial lung diseases and lung carcinoma. The datasets of the lung diseases considered in this study are the large cell lung carcinoma and small cell lung carcinoma. Lung cancer or lung carcinoma is currently the most frequently diagnosed major cancer and the most common cause of cancer mortality in males worldwide. This is largely due to the effects of cigarette smoke.

An international system of tumor classification is important for consistency in patient treatments and to provide the basis of epidemiological and biological studies. In developing this classification, pathologists have tried to adhere to the principles of reproducibility, clinical significance and simplicity, and to minimize the number of unclassifiable lesions. Most of this classification is based on the histological characteristics of tumors seen in surgical or needle biopsy, and is primarily based on light microscopy, although immune histochemistry and electron microscopy findings are provided when necessary. The methodology used in this work defined that the images are pre-processed for the removal of the noises and contrast enhancement is done for obtaining the enhanced images. Feature extraction is frequently used as a preprocessing step to machine learning where the Gabor filter is used in texture analysis. The feature selection method such as the Information Gain, correlation based feature selection, Principal Component Analysis with optimization of the genetic algorithm are done. The feature outputs are combined by watershed segmentation and the fuzzy C means clustering combines the data that belongs to two or more clusters. The Naive Bayes classifier is used to classify the images and the results are shown with the performance measures.

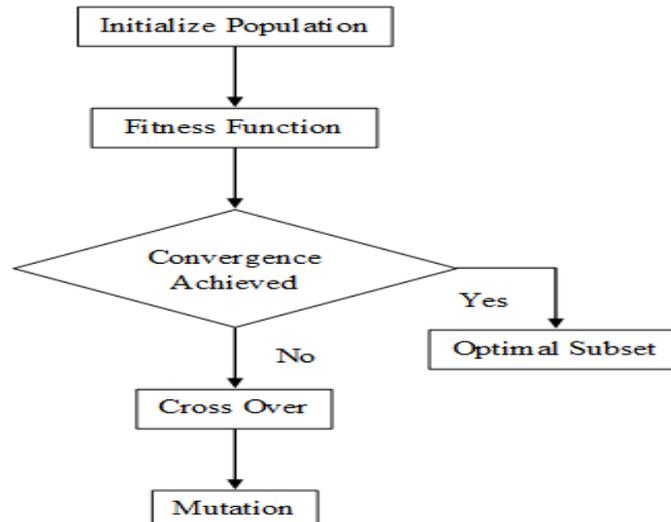
II.SYSTEM MODEL

The ROI region in the lung CT images is obtained and the features were extracted from those regions. Four different features were extracted from the images. B-HOG features, Wavelet features, LBP features and CVH features. B-HOG and LBP were texture based features while Wavelet and CVH were spatial features. The combination of the spectral and the spatial features were more effective in the identification of the diseases in the images. 18 B-HOG features, 26 wavelet features, 96 LBP features, and 40 CVH features were extracted and hence totally 180 features. The extracted features were very large in number and hence the best features were selected from the extracted features. For the selection of the best features fisher criterion based genetic optimization is employed. To find the histogram, small patches of the same size as in the unsupervised step are extracted by sliding a window over each training image in a class. These patches are then converted to the appropriate representation as used in the previous step. Finally, a histogram of textons is computed for the image by comparing each and every patch representation in that image with all textons in the dictionary using a similarity measure to find the closest match and updating the corresponding histogram bin based on the closest match found.

The overall process includes Feature extraction, Feature selection, Classification and Performance analysis. The image segmentation algorithm based on the Fisher criterion function and developed genetic algorithm is introduced. The simulation results show that the new algorithm not only segment the image satisfactorily with strong anti-noise ability, but also run efficiently. The developed genetic algorithm can find the global optimum solution rather than the local solution. The algorithm is one kind of practical, effective segmentation algorithm. In the classification stage, to classify a test image, the same steps as in the learning stage are followed to find the features for the test image. This includes extraction of small patches from each test image in a class, converting the patches to the appropriate representation, finding the closest match to these patches from the dictionary, and computing the normalized histogram of obtained closest textons to define a feature vector for the image. In most of GA-based feature selection methods, the feature selection result is represented by a binary string. Each bit in the string corresponds to a feature, where the value 1 indicates that the feature is selected and 0 indicates that the feature is discarded. Different from these methods, assign a weight to each feature and evolve the weights. It is more reasonable and more accurate for measuring the importance degree of a feature than the hard value of 0 or 1.

III. FITNESS FUNCTION USING FISHER CRITERION

The Fisher criterion measures the distance among all the classes and the divergence within the members of each class. It reflects the classification accuracy under the absence of classifiers. The proposed FIG algorithm flowchart is shown below:



The fitness of the individuals is computed as follows : Let w be the individual required to be evaluated. Let $X^{i,j} = (x_1^{i,j}, x_2^{i,j}, \dots, x_d^{i,j})$ be the full vector of the j th example of the i th class. The first step is to calculate the mean of feature vectors belonging to the i th class as where ' n_i ' is the number of examples of the i th class and the mean of feature vectors of all the training examples is computed as:

$$m = \sum_{i=1}^c \sum_{j=1}^{n_i} X^{i,j} / \sum_{i=1}^c n_i \quad (1)$$

The corresponding class mean is computed as:

$$S_w = \sum_{i=1}^c \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{k=1}^d W_k (x_k^{i,j} - m_k^i)^2 \quad (2)$$

the weighted distance between the classes is computed as:

$$S_B = \sum_{i=1}^c \sum_{k=1}^d W_k (m_k^i - m_k)^2 \quad (3)$$

Finally, the Fisher criterion can be calculated as maximizing S_B and minimizing S_w simultaneously. Thus the fitness function for finding 'w' is:

$$f(w) = S_w / S_B \quad (4)$$

The optimal 'w' is taken as the one that minimizes the equation. Here, k-NN classifier is used to obtain data-driven threshold. Nine thresholds from 0.1 to 0.9 are used to select the features. The feature subset leading to the best CAR is taken as the final selection result and the corresponding threshold as the optimal one.

IV. RESULT AND DISCUSSION

In the fig 1, the extracted region of interest which represents the affected region is shown

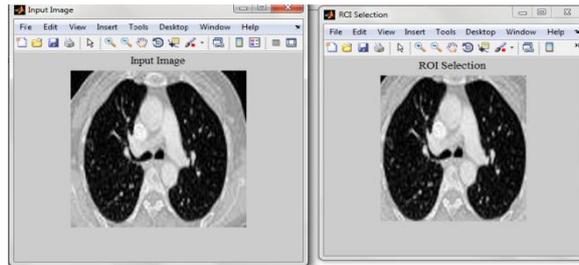


Fig. 1 Region of interest (ROI) of Lung image

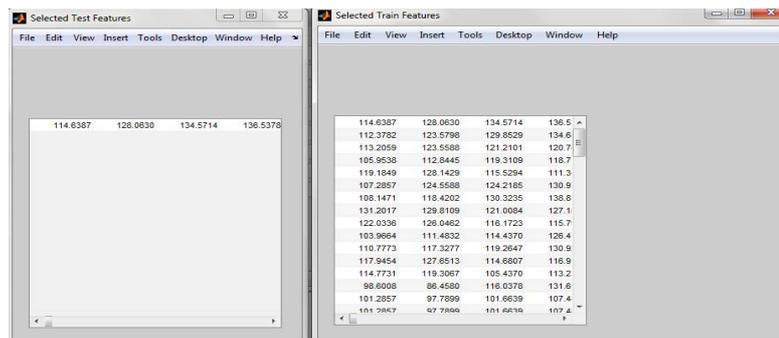


Fig .2 Selected train and test features

In Fig 2, it shows the selected test and train features for further identification of lung diseases. Each function is provided with two execution modes: Training and Testing. The function is called with the training data to obtain the corresponding classifier. Then it is evaluated on the test data by calling the function with the testing mode. The Fisher criterion is applied to evaluate feature selection results, based on which a genetic optimization algorithm is developed.

In Fig 3, the best fitness calculated after the feature selection process is 6.814 combined with the genetic algorithm.

```

4
8
16
32
8.4451e-115

Optimization terminated: average change in the fitness value less than options.TolFun
-----
Iteration = 1 ; Crossover Population = 0.19461 ; Mutation Population = 0.19461
-----
Iteration = 2 ; Crossover Population = 0.04734 ; Mutation Population = 0.04734
-----
Iteration = 3 ; Crossover Population = 0.01152 ; Mutation Population = 0.01152
-----
Iteration = 4 ; Crossover Population = 0.00280 ; Mutation Population = 0.00280
-----
Iteration = 5 ; Crossover Population = 0.00068 ; Mutation Population = 0.00068
-----
The Algorithm Converges At The Iteration = 5
-----
Fitness Value = 6.814

```

Fig .3 Calculated fitness function

In fig 4, Disease Grand grass opacity (GG0) is identified

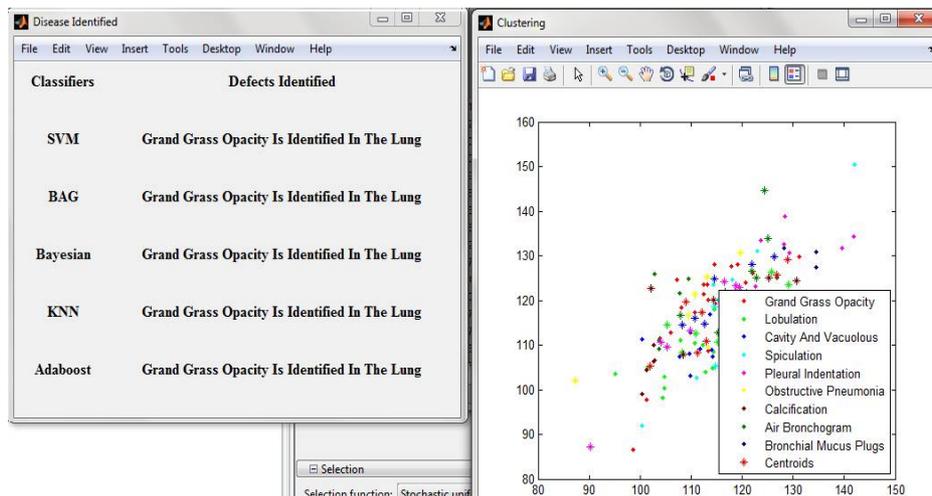


Fig .4 Disease identified

In fig 5, the features extracted from real lung CT images showed that the cooperation of FIG and SVM achieved the average SE of 100%, the average SP of 98.7%, and the classification Accuracy rate of 98.8%. The performance of the process were measured based on the performance metrics like Accuracy, Sensitivity and Specificity. The results show that while aggregation of single decisions by SVMs over various k values using multiple classifier systems helps to improve the results compared to single SVMs.

	Accuracy(%)	Sensitivity(%)	Specificity(%)
SVM	98.8889	100	98.75
BAG	92.2222	100	91.25
Bayesian	92.2222	100	92.50
KNN	88.8889	100	87.50
Adaboost	78.8889	100	87.50

Fig .5 Performance Measures

V.CONCLUSION

CT Lung images were taken as the input. The dataset consists of lung images with nine different type of diseases. The ROI were selected from the CT lung images because the other regions may contain some unwanted information's. For the extraction of the four different types of features were extracted. For the selection of the best features from the extracted features genetic algorithm with fisher criterion is employed. The selected features were then classified using five different classifiers in order to find the disease in the lung images. The performance of the process is measured based on the performance like Accuracy, Sensitivity and Specificity of the classifiers. The performance measured indicates that SVM classifier is efficient using the selected best features. The process can be further enhanced by combining the fisher criterion with other feature selection methodologies. The selection of the best features were based on the application of the different objective functions in the feature selection process.



REFERENCES

1. Brodley .D , A. Kak, L. S. Broderick, ‘ Unsupervised feature selection applied to content-based retrieval of Lung images ’, vol. 25, no. 3, pp 373–378, dec(2013)
2. Chen .G, Y. Xu, Y. Ma, ‘ Neural network ensemble-based computer-aided diagnosis for differentiation of lung nodules on CT im-ages clinical evaluation ’, Acad. Radiol. , vol. 17, no. 5, pp. 595–602, april (2012)
3. Chapman B.E, and B. Zheng, ‘A multistage approach to improve performance of computer-aided detection of pulmonary embolism depicted on CT images: preliminary investigation’, vol. 58, no. 6, pp. 1519–1527, june (2011)
4. Dang, K. Pang, and Y. Zhao, ‘ A novel approach of computer-aided detection of focal ground-glass opacity in 2D lung CT images’, vol. 8670, pp. 86702W, 1–6, july (2012)
5. Depeursinge .O , J. Iavindrasana, and A. Hidki, ‘A classification framework for lung tissue categorization’, vol. 6919, pp. 69190C, 1–12, may(2008)
6. Gangeh Y, L. Sorensen, S. B. Shaker, M. S. Kamel, and M. D. Bruijne, ‘ Multiple classifier system in texton-based approach for the classification of CT images of lung ’, vol. 6533, pp. 153–16, dec(2011)
7. Goldberg and K. Deb, ‘ A comparative analysis of selection schemes used in genetic algorithms ’, Urbana, vol. 51, pp. 61801–2996, aug (2012)