Lung cancer detection using CXRs and pattern recognition

Author A[1], Author B[1], Author C[1], Author D[1] and Author E[1]

[1]Department of Computer Science, LaTeX University

Abstract

Cancer is a leading cause of death worldwide. Lung cancer is a type of cancer that is considered as one of the most leading causes of death globally. This paper aims to elaborate difference between 4 strategies based on 3 different classifiers (SVM & KNN & XGBOOST) and Wavelet transformation by it's all families as a Feature Extraction. and Wavelet transformation by it's all families as a Feature Extraction.

1 Introduction

Lung cancer is one of the most serious cancers in the world. Survival from lung cancer is directly related to its growth at its detection. The earlier the detection is, the higher the chances of successful treatment are. Chest X-ray image has been used for detecting lung cancer for a long time. The early detection and diagnosis of pulmonary nodules in chest X-ray image are among the most challenging clinical tasks performed by radiologists.

Many techniques have been developed in order to increase the detection accuracy rates of lung cancer in CAD systems. This comparative study will elaborate difference between 4 strategies based on 3 different classifiers (SVM & KNN & XGBOOST) and Wavelet transformation by it's all families as a Feature Extraction.

1. classify CXRs without lung extraction
2. classify CXRs with lung extraction
3. Nodule Detection in CXRs without lung extraction
4. Nodule Detection in CXRs with lung extraction

2 Related Work

Many techniques have been developed in order to increase the detection accuracy rates of lung cancer in CAD systems. For instance, Hardie et al. [1] presented a CAD system for pulmonary nodule detection in chest radiography. The proposed system was tested using a data set that consists of 167 chest radiography that contains 181 lung nodules; the system utilized adaptive distance-based threshold algorithm for nodule segmentation, after that, features were computed for each nodule using geometric features, intensity features and gradient features. Lastly, a Fisher linear discriminant classifier was used to c1assity the computed features. The system could detect 78.1 % of those nodules. Lee et al. [2] proposed a lung nodule detection using an ensemble classifier aided by clustering. Lung scans of 32 patients who included 5721 images were used to test the method. The system obtained sensitivity of 98.33% and specificity of 97.11 %.An automatic method for lung cancer detection was also introduced by Sousa et al. [3]. The system has six stages, namely, thorax extraction, lung extraction, lung reconstruction, structure's extraction; tubular structures elimination, and false-positive reduction. Each stage performs a specific task that resulted in detecting lung nodules. The sensitivity achieved was 84.84%, and specificity achieved was 96.15%. These were examples of systems that have been

developed to detect lung cancer. General issues of existing systems are the high number of false positives and false negatives. Therefore, there is a continuous need and importance to develop computer-aided diagnosis to help in lung cancer detection and prediction. According To resources mentioned in references The most technique used to extract lung field is Active shape model (ASM), and the most classifier used is KNN and SVM, So this work aims to do comparative study based on three different classifier KNN, SVM , XGBoost.

3 Materials

The proposed systems was tested using JSRT Database includes 154 abnormal chest radiographs, each with a solitary pulmonary nodule, and 93 non-nodule chest radiographs. These original screen-film images were digitized with a 0.175-mm pixel size, matrix size of 2048 2048, and 12 bits of gray scale.

4 Method

4.1 Feature Extraction

Feature Extraction based on Wavelet Transform Feature extraction is an important step in CAD systems. The extraction of features that represent medical images of specific organs is an important issue. Multi-resolution methods such as Wavelet Transform (WT) have been widely used in feature extraction due to its quality compared with other feature extraction methods. In PYHTON, wavelet is implemented as a library called pywavelets.its offers many wavelet functions that can be utilized. Wavelet families such as Haar (haar), Daubechies (db), Symlets (sym), Coiflets (coif), Biorthogonal (bior), Reverse biorthogonal (rbio), "Discrete" FIR approximation of Meyer wavelet (dmey), Gaussian wavelets (gaus), Mexican hat wavelet (mexh), Morlet wavelet (morl), Complex Gaussian wavelets (cgau), Shannon wavelets (shan), Frequency B-Spline wavelets (fbsp), Complex Morlet wavelets (cmor), with scales(Levels) from 1 to 6 level, with (ad,da,dd) horizental,vertcal,diagonal detailes coefficients. such as family(haar) , level(4) that return (ad,da,dd) detailes coefficients and one of them is classified

4.2 Feature Selection

Once the feature extraction stage has been executed, a huge number of coefficients will be produced. It is important to reduce the coefficients by selecting those coefficients that contains the most important information that would contribute to high accuracy and ignoring the remaining. For this reason,we use (low variance) to remove these coefficients
VarianceThreshold is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features, i.e. features that have the same value in all samples.

4.3 Classification

python offers Simple and efficient tools for data mining and data analysis Accessible to everybody, and reusable in various contexts Built on NumPy, SciPy, and matplotlib Open source, commercially usable - BSD license called scikit-learn
Classification methods used in proposed systems are support vector machine(SVM),K-Nearest Neighbor (KNN) ,xgboost classifier.
dataset are dived to 80% training set and 20% testing set then we shuffle data and repeat that process of classification 50 times and get average result

5 Experimental Study

5.1 classify CXRs without lung extraction (system 1)

5.1.1 Feature Extraction

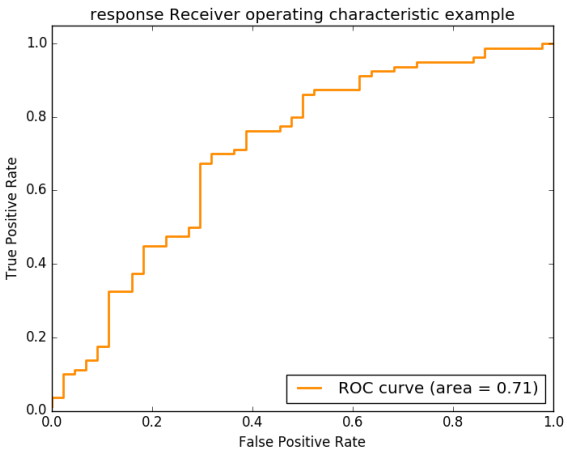
Feature Extraction based on Wavelet Transform with families 'haar', 'db', 'sym', 'coif', 'bior', 'rbio', 'dmey', 'gaus', 'mexh', 'morl', 'cgau', 'shan', 'fbsp', 'cmor', and scales(Levels) 4,5,6

5.1.2 Classification

We dived the dataset to 80% training set and 20% testing set then we shuffle data and repeat

that process 50 times and get average result we do that with 3 different classifier KNN, SVM and XGBoost

5.1.3 ROC



5.2 classify CXRs with lung extraction (system 2)

5.2.1 Preprocessing

To eliminate false positives that might occur outside of the lung, it is important to obtain a very accurate segmentation of the lung boundaries. In this work we used ASM(active shape model) machine learning technique with dataset lung field landmarks to extract lung from image

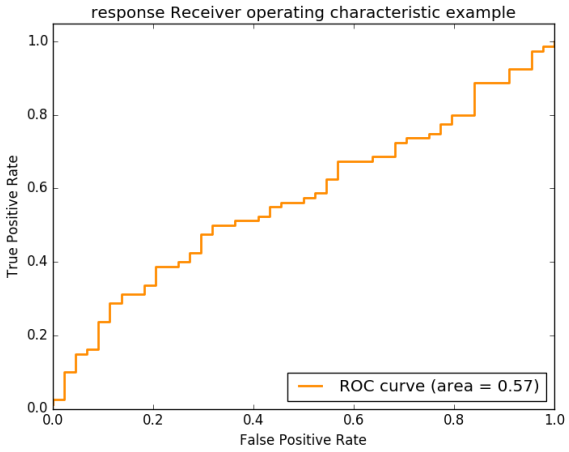
5.2.2 Feature Extraction

Feature Extraction based on Wavelet Transform with families 'haar', 'db', 'sym', 'coif', 'bior', 'rbio', 'dmey', 'gaus', 'mexh', 'morl', 'cgau', 'shan', 'fbsp', 'cmor', and scales(Levels) 4,5,6

5.2.3 Classification

We dived the dataset to 80% training set and 20% testing set then we shuffle data and repeat that process 50 times and get average result we do that with 3 different classifier KNN, SVM and XGBoost

5.2.4 ROC



5.3 Nodule Detection in CXRs without lung extraction(system 3)

5.3.1 Preprocessing

In this expirement we divide our images to patches each patch equal to 64*64 pixel
The abnormal regions were selected based on the coordinates that are provided with the dataset, and the normal regions were extracted randomly from the 93 non nodule images for each abnormal region we get 7 different orientation

1- orignal postion
2- 90 degree rotation
3- 180 degree rotation
4- 270 degree rotation
5- flip horizontal
6- flip horizental with 90 degree rotation
6- flip horizontal with 180 degree rotation
7- flip horizontal with 270 degree rotation

In JSRT there exist 156 noduled image from each image we get 7 different patches, so we will have 1078 noduled patches

Then we divide normal images into equal patches (64*64 pixel) In JSRT there exist 93 normal image, each image is divided into 1024 patch so we will have 95,232 normal patch

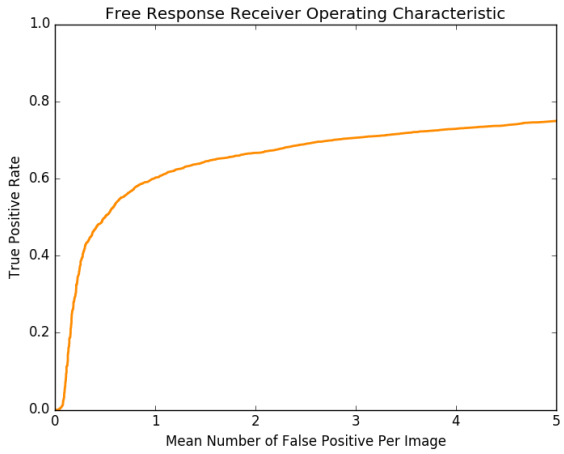
5.3.2 Feature Extraction

Feature Extraction based on Wavelet Transform with families 'haar', 'db', 'sym', 'coif', 'bior', 'rbio', 'dmey', 'gaus', 'mexh', 'morl', 'cgau', 'shan', 'fbsp', 'cmor', and scales(Levels) 1,2,3

5.3.3 Classification

At first we choose 1078 random sample from normal patches and take all noduled patches then we divide them to 80% training set and 20% testing set then we shuffle data and repeat that process 50 times and get average result we do that with 3 different classifier KNN, SVM and XGBoost

5.3.4 FROC



5.4 Nodule Detection in CXRs with lung extraction (system 4)

5.4.1 Preprocessing

We use Active Shape Model to extract lung field from CXRs and divide our images to patches each patch equal to 64*64 pixel

n this expirement we divide our images to patches each patch equal to 64*64 pixel
The abnormal regions were selected based on the coordinates that are provided with the dataset, and the normal regions were extracted randomly from the 93 non nodule images for each abnormal region we get 7 different orientation

1- orignal postion
2- 90 degree rotation
3- 180 degree rotation
4- 270 degree rotation
5- flip horizontal
6- flip horizental with 90 degree rotation
6- flip horizontal with 180 degree rotation
7- flip horizontal with 270 degree rotation

In JSRT there exist 156 noduled image from each image we get 7 different patches so we will have 1078 noduled patches

Then we divide normal images into equal patches (64*64 pixel) In JSRT there exist 93 normal image, each image is divided into 1024 patch and remove patches that are outside the lung

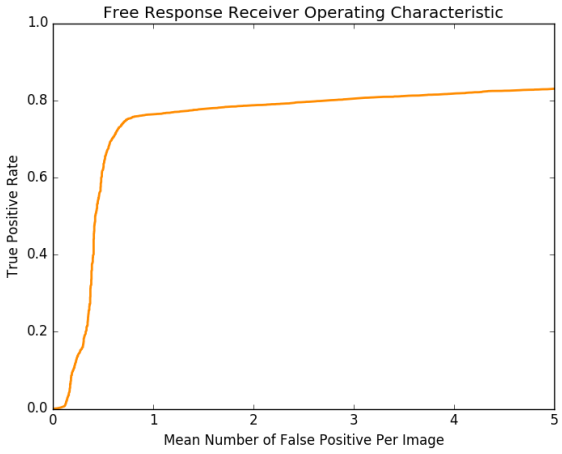
5.4.2 Feature Extraction

Feature Extraction based on Wavelet Transform with families 'haar', 'db', 'sym', 'coif', 'bior', 'rbio', 'dmey', 'gaus', 'mexh', 'morl', 'cgau', 'shan', 'fbsp', 'cmor', and scales(Levels) 1,2,3

5.4.3 Classification

At first we choose 1078 random sample from normal patches and take all noduled patches then we divide them to 80% training set and 20% testing set then we shuffle data and repeat that process 50 times and get average result we do that with 3 different classifier KNN, SVM and XGBoost

5.4.4 FROC



6 Result

The table shown in best results achieved in following table :

	SVM	KNN	XGBOOST
System 1	74.44%	84.22%	96.0%
System 2	83.0%	83.8%	82.6%
System 3	82.25%	93.14%	92.13%
System 4	90.22%	86.34%	99.98%

Best result achieved in system 1 with SVM used db32 level 5 (ad) as feature extraction and with KNN used coif4 level 5 (ad) as feature extraction and with XGBOOST used bior5.5 level 4 (ad) as feature extraction

Best result achieved in system 2 with SVM used sym6 level 4(ad) as feature extraction and with KNN used sym6 level 4(ad) as feature extraction and with XGBOOST used sym6 level 4(ad) as feature extraction

Best result achieved in system 3 with SVM , bior3.9 used level 1 (ad) as feature extraction and with KNN , bior2.2 used level 2 (ad) as feature extraction and with XGBOOST used db3 level 3 (ad) as feature extraction

Best result achieved in system 4 with SVM used db7 , level 2 (ad) as feature extraction and with KNN used rbio3.1 level 3 (ad) as feature extraction and with XGBOOSt used sym4 level 2 (dd) as feature extraction

7 Conclusion

After trying different expirements we can conclude that:

- System with lung field extraction and segmented patches get the best accuracy compared with others systems

- XGboost has better results with wavelet Transformation

8 References

1. Russell C. Hardie, Steven K. Rogers, Terry Wilson, Adam Rogers; Performance analysis of a new computer aided detection system for identiJYing lung nodules on chest radiographs. Medical Image Analysis 12 (200S); pp. 240-25S.

2. S.L.A. Lee, A.Z. Kouzani, EJ. Hu, Random forest based lung nodule classification aided by clustering. Computerized Medical Imaging and Graphics 34 (20 I 0); pp. 535-542

3. Joao Rodrigo Ferreira da Silva Sousa, Arist6fanes CorrVea Silva, Anselmo Cardoso de Paiva, Rodolfo Acatauassu Nunes, Methodology for automatic detection of lung nodules in computerized tomography images. Computer Methods and Programs in Biomedicine 9S (2010); pp. I-14.