ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse97952020.pdf
https://doi.org/10.30534/ijatcse/2020/97952020

Lung cancer detection with prediction employing machine learning algorithms

Krishna Prasad SJ[1], Aneesha Johnson[2], Mohana Kumar S[3]
[1]Department of Electronics and Telecommunication, M S Ramaiah Institute of Technology, Bangalore, India, krishnaprasad@msrit.edu
[2]Research scholar, Department of Electronics and Telecommunication, M S Ramaiah Institute of Technology, Bangalore, India, aneesha@scpc.edu.in
[3] Department of Computer Science, M S Ramaiah Institute of Technology, Bangalore, India, mohanks@msrit.edu
*Corresponding author krishnaprasad@msrit.edu

ABSTRACT

Lung cancer-related deaths are increasing globally every year. The prediction of the onset of cancer in patients can help doctors in the process of decision making regarding their medications and treatments. This paper proposed a novel system which is capable of detecting and predicting the presence of cancerous nodules in the patient's lungs. The proposed system employs machine learning algorithms like support vector machine (SVM) and deep learning algorithm like convolutional neural network (CNN), to perform the classification, using an extensive lung cancer repository database namely, the UCI repository. In the first stage of cleaning of images, they are pre-processed and further post-processed. The pre-processing step includes RGB to grey scale conversion and further removal of noise is achieved by using Non-Local Means (NLM) filter. In the second stage of development image segmentation has been realized using Otsu's algorithm and feature extraction has been realized with the help of Grey Level Co-occurrence Matrix (GLCM). Finally, lung cancerous images are classified using the two classifiers and the accuracy of their classifications are being compared and tabulated.
.

Key words : Computed Tomography, Grey Level Co-occurrence Matrix Non local mean, Otsu's thresholding algorithm.

1. INTRODUCTION

Cancer is one of the major diseases the world population is battling today, which is found across people irrespective of nationality, gender and communities. In cancer, the cells begin to divide without stopping and spreads to the surrounding tissues. Many cancers form solid structures called tumors, which are huge mass of tissue. As per statistics available there are 2.09 million fresh lung cancer subjects detected every year. Also among 1.76 million people deaths due to cancer every year lung cancer related death contribution is major as per records of World health organization. Further lung cancer is second most cancer affecting among population worldwide. Though the diagnosis of lung cancer is undertaken by lung biopsy easy way is to investigate sample of CT image of lung cells in the laboratory for any possible visualization of lumps or nodules in the images. A noninvasive way happens to be computed tomography (CT) scan which captures the lung tumors present in the patients lungs. It will also capture the shape, size and position of lung tumors and can find the enlarged lymph nodes that contains cancerous cells that has spread. This CT image investigation can also be used for capturing masses in the liver, brain, adrenal glands and other organs that might have developed due to the onset of lung cancer. Lung cancer related CT images are chosen as an experimental database because there is presence more information compared to X-rays as they have been synthesized images captured from multiple angle trajectories. In this work, the detection and classification of the cancerous and noncancerous lung nodules is done through an semi- automated system using image processing algorithms for segmentation and feature extraction . Further, machine learning algorithm have been developed for classification namely, Support Vector Machine (SVM). Comparisons of results are also being made by developing deep learning algorithm namely, convolutional neural network (CNN) .Comparison and accuracy of result of the performance parameters of both of algorithms have been documented.

The work has been arranged as follows. Section 2 documents literature review of previous related works on cancer detection and prediction. Section 3 documents the methodology of work undertaken. Section 4 documents the simulated results obtained from both algorithms. Further, the proposed system makes a comparison between the two classifiers and their performance parameters are tabulated.

2. PREVIOUS WORKS

In one of the work [1] an algorithm called as entropy degradation method (EDM) has been proposed, which is employed to detect the small cell lung cancer (SCLC) from the CT scans. It is observed that the developed algorithm achieves an accuracy of 77.8%. In next related work [2] a fifteen layer two dimensional DCNN network architecture is proposed, which is used to classify the pulmonary nodule. This work proposed does not include the segmentation step in the processing stage to extract the patch. Therefore by involving this step in the data pre-processing, chances of improvement in accuracy of the proposed system are possible. In further related work [3] various optimization algorithms namely, k nearest neighbor (KNN), support vector machine, method have been evaluated to detect the tumor in the lungs. It is observed that on using more number of optimization algorithms, improvements in accuracy of the system are possible .In next related work [4] a system has been developed using a generative adversarial network (GAN) for boosting the performance of the classifier and the accuracy achieved is around 94%. In next related work [5] an automated model has been developed to classify the abnormalities in mammogram. After processing the image, it employs a KNN classifier for the early detection of breast cancer. It is observed that the developed algorithm achieves an accuracy of 92%. In yet another work [6] the focusis to extract the lung region and the desired features using grey level co-occurrence matrix (GLCM). The drawback of this work is that a classifier has not been employed to classify the given images.In another related work [7] a model is developed using ANN algorithm with back propagation for training the developed model. This work employs several processing steps to enhance the image. And the accuracy achieved is 80%. In yet another work [8]a lung classification model is developed using SVM classifier. In this work the input images are taken from an online repository called kaggle. To remove the noise and preserve edges the system employs a median filter. Watershed segmentation is the type of segmentation technique used and GLCM is used for feature objects to be extracted. Finally the classification is done using a SVM classifier and it is observed that an accuracy of 96.7% is achieved. In another work [9] a supervised and unsupervised approach is being adapted to improve the characterization of lung and pancreas tumor. It employs a deep learning algorithm like 3D convolutional neural network (CNN) and a machine learning algorithm namely support vector machine (SVM). In another related work [10][14] a computer aided detection (CAD) model is being developed. The proposed work comprises of four stages which include, noise removal using median filter, performs morphological operation for image enhancement, employs an adaptive thresholding algorithm, feature extraction and classification using k-nearest neighbor (KNN). The developed model has achieved a detection rate of 90%. In yet another work [11algorithm based on convolution neural network has been realized .Perspective of this work is pipelining technique. Focus of this work is to tradeoff between delay in time and consumption of power to get overall optimal performance of the network.In another work [12] a novel computer aided detection (CAD) system has been proposed for pulmonary nodules using multi-view convolution networks (ConvNets). The input data is collected from a public database repository namely LIDC-IDRI dataset. The proposed technique touches large values of sensitivities of detection 90.1%.and 85.4% . In another related work [13] a pulmonary nodule detection method is being developed. In the proposed work the extracted features are applied to the support vector machine (SVM) to classify the nodule candidate into nodule and non-nodule. Also the proposed work achieved a sensitivity of 95.28%.

3. METHODOLOGY

Lung cancer detection system faces many challenges, from works investigated it is seen that various techniques have been proposed to detect the presence of tumors in the lungs. UCI repository has been used in many works for process of image acquisition. For the process of de-noising the images, filters namely median filter, is being used. Most of the systems employ image processing steps like image enhancement, image segmentation and feature extraction. The image segmentation is done using Otsu's thresholding algorithm and the features are extracted using the (GLCM). It is observed that the support vector machine (SVM) and convolutional neural network (CNN) outperforms the other types of classifiers. Hence in our proposed work both the classifiers have been used and also their results are being compared to identify the better classification performance.

The proposed work has been realized in four stages for lung cancer classification. During stage one data acquisition has been from most popular repository used by researchers of cancer namely the University of California Irvin(UCI) CT lung cancer image repository. At the level of second stage acquired lung images are to grey scale images from RGB images. Also noise removal is achieved by employing NLM filter. During the third stage images are subjected to segmentation using Otsu's thresholding and features are extracted using GLCM (Gray Level Co- occurrence Matrix). More number of features are used to increase the performance of the classifier. In stage four, for the classification purposes two classifiers namely SVM (Support Vector Machine) and CNN (Convolutional Neural Network) classifiers are employed . Further, accuracies of

both the classifiers are compared. The Figure 1 represents the block diagram of the proposed work.

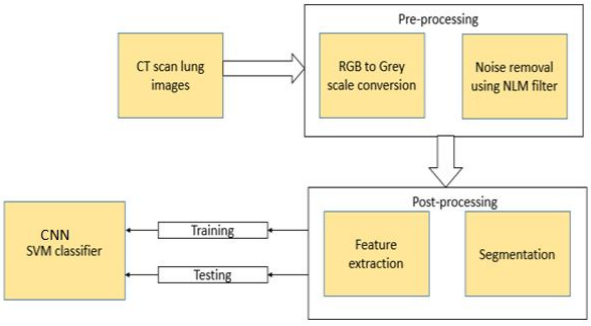

Figure 1: Block Diagram of the model

3.1 Image Aquisition

UCI database repository being an extensive repository of CT(computed tomography) images of lung cancer has been used in this work for testing cancerous images of patients. These images are in Diacom format. The CT scan images are synthesized images with high resolution and less noise.

3.2 Preprocessing of images

CT images acquired from UCI repository is subjected to further conversion process. Since repository images are in RGB format they are transformed to grayscale images .Main objective for conversion is to facilitate easy processing. Predominant characteristic of medical images being that they are low by contrast and during acquisition process random noises get added to them. Hence, removal and cleansing these noises is mandatory especially in analysis of medical images. NLM filter provides a powerful framework in which the robustness of de-noising the image is high.

Given a noisy discrete image is

$$v = \{v(i) \mid i \in I\} \quad (1)$$

The estimator of the noise *image(v)* at any *location (i)* is given by NL[v](i).

$$NL[v](i) = \sum_{j \epsilon I} w(i,j)v(j) \quad (2)$$

Where the weights family {w(i,j)}, depends on the alikeness between the pixcels *i* and *j* and suit the conditions,

$0 \leq w(i,j)v(j) and\ \Sigma\, w(i,j) = 1.$

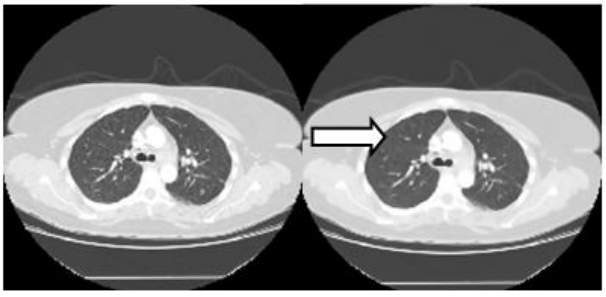

Figure 2: Filtered CT scan lung image

3.3 Image Segmentation

One of the most important processing step to be followed is said to be segmentation. Here the image is portioned into multiple segments. This work employs the Otsu's thresholding for segmentation of lung images. Here the input de-noised grayscale image is first converted to a binary image and then masked with the extracted patch to obtain the region of interest (ROI). Every image possess two types of pixels namely, foreground and background pixels.

Here the weighted sum of variances within two classes is defined as:

$$\sigma_w^2(t) = w_0(t)\,\sigma_0^2(t) + w_1(t)\,\sigma_1^2(t) \quad (3)$$

Where the 'w_0' and 'w_1' are the probabilities of the two classes separated by the threshold, 't', ${\sigma_0}^2$ and ${\sigma_1}^2$ are the variences of these two classes.

Hence the total variance is defined as follows:

$$\sigma^2 = \sigma_w^2(t) + w_0(t)[1 - w_0(t)][\mu_0(t) - \mu_1(t)]^2 \quad (4)$$

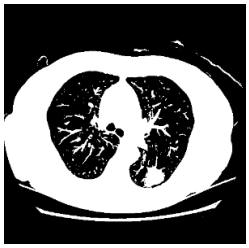

Figure 3: The segmented lung region using Otsu's thresholding

3.4 Extraction of Features

Researchers in medical image processing have deployed various quality techniques for extraction of features among them one of the quality technique is the use of GLCM(Grey Level Co-occurrence matrix). This technique has been popular technique for extraction of texture features from preprocessed image. Here the pixel values are extracted from the segmented patch. The co-occurrence matrix shall be developed by using the NE 1px method that is moving one pixel in the north east direction. On developing the co-occurrence matrix, the features can be extracted from the developed matrix [5]. A total of 7 different features are extracted namely, area, mean, standard deviation, entropy, variance, RMS and smoothness. The terms are defined as follows:

Energy: Sum of the squares of all elements in GLCM matrix where *i* and *j* are row and column of matrix respectively. It also gives an intuitive idea of homogeneity of pixels. Its value ranges from 0 to 1. Where the *p* denotes the probability of pixels.

$$\text{Energy} = \sum_{i,j} p(i,j)^2 \quad (5)$$

Entropy: Defines of variations of pixels values of an image which is statistical in nature.

$$\text{Entropy} = -\sum(p.* log2(p)) \quad (6)$$

Area: The total number of white pixels in the region of interest (ROI) of image is defined by this feature.

Mean(μ)**:** It is defined as average of all the values of pixels in region of interest (ROI) of image of interest.

$$\mu = \sum_{i=1}^{M}\sum_{J=1}^{N}\frac{p(i,j)}{MN} \quad (7)$$

Standard Deviation: Statistic of average distance between pixels values and mean is defined by this feature

$$\boldsymbol{\sigma} = \sqrt{\sum_{i=1}^{M}\sum_{J=1}^{N}\frac{(p(i,j)-\mu)^2}{MN}} \quad (8)$$

Smoothness: The measure of contrast of grey level is defined as the relative smoothness.

$$R=1-\frac{1}{1+\sigma^2} \quad (9)$$

Root mean square (RMS): The arithmetic mean Y of the squares of the mean values is defined as Root mean square value of the image

$$Y=\frac{\sqrt{\sum_{i=1}^{M}|\mu|_{ij}^2}}{M} \quad (10)$$

Notations in the from equation (5) to equation (10), are defined as follows.

i,j – represent two intensities

p(i,j) – intensity of the pixel at that point

MN – denotes the size of the image as *M by N*

μ - represents the mean

σ – represents the variance

R- relative smoothness

3.5 Classification

Support Vector Machine (SVM) is a machine learning algorithm used for classification. In this machine learning based algorithm, n features are required to identify a point in n dimension feature space [8]. Features n are calculated from GLCM matrix. Further decision boundaries which separate different classes are hyper planes as highlighted in Figure 4.

The equation that defines the decision surface which separates the classes is a hyper plane is given as follows:

$$w^T x + b = 0 \quad (11)$$

w is a vector related to weight
x is a vector related to input
b is a bias related constant

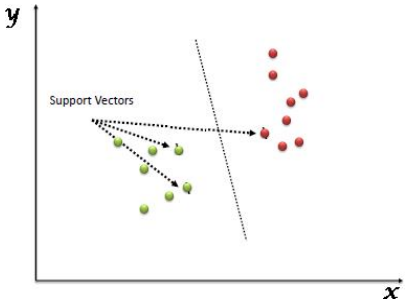

Figure 4: A hyper-plane differentiating two classes of data points

The convolution neural network (CNN) uses the minimal pre-processing steps in the image classification algorithm. The convolution neural network has different layers namely input layer, convolution layer, max pooling layer, RELU layer, fully connected layer and the output layer [4]. The Figure 5 shows a multi layer convolution neural network.

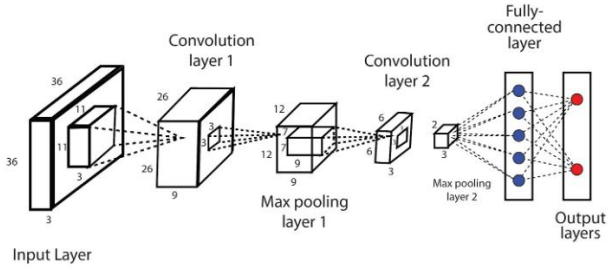

Figure 5: A two layer convolution neural network

4. RESULT AND DISCUSSION

For the classification of cancerous and non-cancerous CT scan lung images, two classifiers are incorporated in the proposed work namely, support vector machine (SVM) and convolutional neural network (CNN). The system is also capable to classify the stage of cancer as final, middle or initial stages depending on the values of features extracted from the GLCM as shown in Table 1. This system is implemented in MATLAB using specific MATLAB functions such as *rgb2gray*, *glcms= graycomatrix(I)* etc., Toolboxes namely image processing toolbox, machine learning toolbox, deep learning toolbox and graphical user interface. The Figure 6-7 represents the GUIDE window showing the final stage detection of cancer using SVM classifier. Similar GUIDE window can be obtained for the CNN classifier. A total of 50 *nos* images were taken as training dataset and 35 *nos* images were used to test the performance of the classifiers. Finally the performance parameters of both the classifiers were obtained by developing the confusion matrix. A Comparison between SVM and CNN classifiers are shown in Table 2. The accuracy obtained in CNN classifier algorithm is higher in comparison with SVM classifier algorithm. It is observed that CNN outperforms SVM classifier with such fast and higher

accuracy algorithm, one can not only detect lung cancer accurately but also can decrease the mortality rate. The SVM classifier gave an accuracy of 78.9% and CNN gave an accuracy of 94.7%.

Table 1: Feature threshold values for different class of classification

SL NO	Features	Stage 1	Stage 2	Stage 3	Normal
1	Area	81	198	362.5	0
2	Mean	0.074	0.131	0.258	0
3	Standard Deviation	4.839	11.831	21.706	0
4	Entropy	245.25	592.66	1086.88	0
5	RMS	0.059	0.059	0.059	0.059
6	Variance	0.285	0.705	1.295	0
7	Smoothness	76.566	184.61	335.67	0

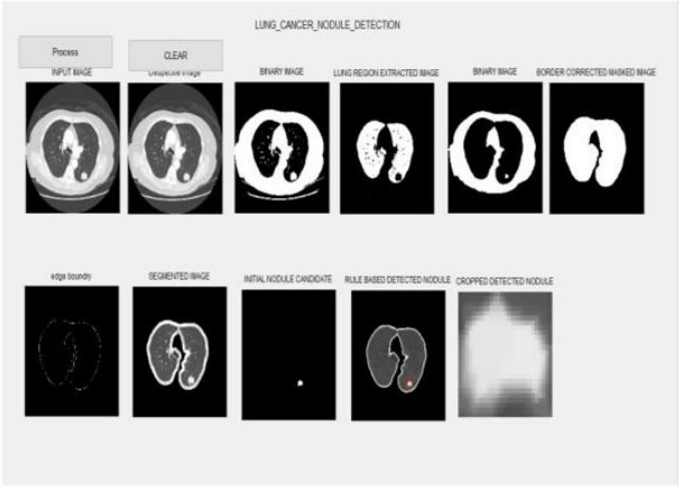

Figure 6: The complete GUIDE window of the output result

Figure 7: Experimental result of classification

Table 2: The performance parameters of the classifiers

Classifier	Accuracy	Sensitivity	Specificity
SVM	78.9%	0.8	0.87
CNN	94.7%	1.0	0.87

5. CONCLUSION

This proposed and successfully implemented machine learning algorithm based classification techniques consists of four stages .In the initial two stages of image acquisition and image preprocessing NLM filter has been deployed for cleansing images corrupted by random noises added during acquisition. In further stages of image segmentation, feature extraction and classification, Otshu's thresholding is used for faster and accurate results, to extract features, Grey level co occurrence matrix has been employed. Finally for classification of four classes, seven features extracted from GLCM matrix have been employed. Both SVM and CNN classifiers are successfully implemented in this work and their performance parameters are documented not only as accuracy but also in confusion matrix. It is observed that CNN outperforms SVM classifier.

ACKNOWLEDGMENT

Authors hereby acknowledge the support of M.S.Ramaiah Institute of Technology, Bangalore for creating ambiance for conducting this study.

REFERENCES

1. Qing Wu and Wenbing Zhao, "Small Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm" in 2017 International Symposium on Computer Science and Intelligent Controls, © 2017IEEE
2. Giang Son Tran, Thi Phuong Nghiem, Van Thi Nguyen, Chi MaiLuong and Jean Christophe Burie, "Improving Accuracy of Lung Nodule Classification Using Deep Learning with Focal Loss", in Journal of Healthcare Engineering, Article ID 5156416, Volume2019
3. K. Senthil Kumar, K. Venkatalakshmi and K. Karthikeyan, "Lung Cancer Detection Using Image Segmentation by means of Various Evolutionary Algorithms", in Computational and Mathematical Methods in Medicine, Article ID 4909846, Volume 2019
4. Onishi, Atsushi Teramoto, Masakazu Tsujimoto, Tetsuya Tsukamoto, Kuniaki Saito, Hiroshi Toyama, Kazuyoshi Imaizumi and Hiroshi Fujita, "Automated Pulmonary Nodule Classification in Computed Tomography Images Using a Deep Convolutional Neural Network Trained by Generative Adversarial Networks", in BioMed Research International, ArticleID 6051939, Volume2019

5. Than Htay and SuMaung, "Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image", in The 18th International Symposium on Communications and Information Technologies, ©2018 IEEE
6. Jaspinder Kaur, Nidhi Garg, Daljeet Kaur, "Segmentation and Feature Extraction of Lung Region for the Early Detection of Lung Tumor", International Journal of Science and Research(IJSR), Volume 3 Issue 6, June 2014
7. Anifah, Rina Harimurti, Haryanto and Zaimah Permatasari, "Cancer Lungs Detection on CT Scan Image Using Artificial Neural Network Back propagation Based Gray Level Co-occurrence Matrices Feature", in ICACSIS, © 2017IEEE
8. R. Ankita, Ch Usha Kumari, MohdJaveed Mehdi, N. Tejashwini, T. Pavani, "Lung Cancer Image- Feature Extraction and Classification using GLCM and SVM Classifier", in International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue- 11, September2019
9. Sarfaraz Hussein, PujanKandel, Candice W. Bolan, Michael B. Wallace, and UlasBagci, "Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches", in IEEE Transactions on Medical Imaging,© 2018IEEE.
10. Pooja Samudre, Prashant Shende, Vishal Jaiswal, "Optimizing Performance of Convolutional Neural Network Using Computing Technique" in5th International Conference forConvergence in Technology (I2CT), ©2019IEEE J.
11. Kuruvilla and K. Gunavathi, "Lung cancer classification using neural networks for CTimages," Computer Methods and Programs in Biomedicine, vol. 113, no. 1, pp. 202–209,2014.

A. Setio, F. Ciompi, G. Litjens et al, "Pulmonary nodule detection in CT images: false positive reduction using Multi view convolutional networks", IEEE Transactions on Medical Imaging, vol.35, no. 5, pp. 1160–1169,2016.

12. Wook-Jin Choi and Tae-Sun Choi, "Automated Pulmonary nodule Detection System in Computed Tomography Images: A Hierarchical Block Classification Approach", Entropy 2013,
13. Bayan Mohammed Saleh et. al. D-Talk: Sign Language Recognition System for People with Disability using Machine Learning and Image Processing IJATCSE, Vol. 9 № 4, 2020, pp. 4374-4382. DOI:https://doi.org/10.30534/ijatcse/2020/29942020