

A Survey on Lung Cancer Detection Techniques

Rohit Bhalerao¹, Harsh Jani², Rachana Gaitonde³, Vinit Raut⁴

Department of Computer Engineering, Mumbai University, MUMBAI

Abstract— The proposed solution uses Image Processing and Deep Learning. Biomedical Image Processing is the latest emerging tool in medical research used for the early detection of cancers. Biomedical image processing techniques can be used in the medical field to diagnose diseases at an early stage. Basically in biomedical Image Processing four steps are involved: Pre-processing, segmentation, feature extraction and classification. Computed Tomography (CT Scans) of lungs of the patients from Lung Image Database Consortium (LIDC) is used as input data for image processing. In pre-processing stage conversion of RGB image to gray-scale image takes place because RGB images are too complex to process. Gray-scale image is further converted to Binary image. In segmentation stage the image is partitioned into segments by watershed segmentation technique. Segmentation stage helps to convert image into a form that is ideal for feature extraction and gives the Region of Interest. After Image Processing, the input images become more efficient and refined. These are input for the Convolution Neural Network. Convolution Filtering, Max Pooling filtering are steps in CNN which train the data to predict whether lung image is cancerous (malignant) or non-cancerous (benign). The proposed system will also take into account the processing power and time delay of the cancer detection process for efficiency.

Keywords: CNN, Deep Learning, LIDC, Image processing, CT scan, Watershed Segmentation.

I. INTRODUCTION

Image Processing is a technique to transform an image into digital form and perform some operations on it, in order to get a more enhanced image and to extract some useful information from it. It includes acquiring input images, noise removal, enhancement and segmentation. In Image Processing, the RGB images will be converted to grayscale and binary. Images will be enhanced and noise will be removed using filters. Blurred effect will be eliminated if any. This improves the quality of input image [13]. The unwanted part of image is cropped using segmentation to obtain the ROI. The Image Processing Toolbox in MATLAB is used to perform the Image Processing stages. Many different algorithms are possible for each stage in image processing [14].

Deep Learning is used for the classification of CT Scan Images as cancerous/non-cancerous. The process of feature extraction in Convolution Neural Networks is such that features are defined and computed by the algorithm itself. During the training stage, input and an output are provided. Based on the given data, the algorithm analyses the features/patterns and for a training data, forms a set of parameters and feature extraction [15]. Based on the computations, the new data can be tested for prediction of a correct output. Convolution Neural Networks consist of an input and an output layer, and multiple hidden layers. The input layers accept inputs and the number of output layers define the number of outputs in the result. Convolution layers are used to define features and parameters. Pooling layers bring together the computations with similar computations. The convolution filter will form a spatially dense output by assigning a common value to a set of matrix pixels. These values decide the output for that image.

II. REVIEW

Computer aided detection/classification systems are gaining popularity in many applications, particularly in the biomedical field. Classification of diseases, infected or uninfected specimen, predictions in cardiology or forecasting in anesthesia, all need artificial intelligence, machine learning and deep learning [16]. CT Scan images provide the best functionality in image classification in cancer detection. Neural networks and SVM machine learning approach show good results. Deep learning CNN is the latest development in artificial intelligence, and is more powerful than machine learning or neural networks. CNN, on proper training models, will provide better results than SVM or ANN for same input data.

III. LITERATURE REVIEW

Computer aided detection of lung cancer have been used over the years in biomedical research and hence there are many systems implemented in this field. The chapter contains literature survey of some papers which are related to the proposed system. Furthermore, the chapter contains an analysis table for all the papers discussed in literature survey.

Following are some papers related to the proposed system:

In 2012, Mokhled S. Al Tarawneh published a comparison paper between different Image Processing techniques and the algorithms they use for a CAD system for lung cancer detection. The main aim of the paper is to detect features for accurate comparisons between images with different processing techniques [1]. Three steps are Image enhancement, segmentation and feature extraction. The aim of image enhancement is to improve quality of image to provide better input for classification. Gabor filter, auto enhancement and fast-fourier transform. Gabor filter has the highest enhancement rate. Image segmentation means to change or simplify representation of image to make is more meaningful for classification and analysis. It includes recognizing the ROI and its separation from image. Thresholding and Watershed methods are used, of which, watershed provides a better quality of segmentation. Feature extraction uses binarization and masking approach. Binarization coins the image of lung as normal or abnormal depending on number of black pixels in image. Masking depends on presence of tumor in ROI. If white masses in image increase, probability of cancer increases. Binarization and Masking, on combined implementation, gives an optimal result.

In December 2017, Suren Makaju, P.W.C Prasad, Abeer Alsadoon and A.K. Singh worked on CAD system of lung cancer with CT Scan Images as primary focus. They believe the that CT Scan Images are the best input data for this research, because it is the best imaging out of MRI, X-Rays or other alternatives [2]. The author studies existing systems to compare the similarities and differences, and makes a current best model on observations. The proposed system is brings improvements to the current best model. The proposed model uses noise removal algorithms before image processing and replaces Gabor filter with a more efficient median filter and gaussian filter. It uses the same segmentation as the current system, i.e., watershed algorithm and promotes a well-defined feature extraction before classification using SVM. The author has used images from LIDC dataset and the system gives a 92% accuracy and 50% specificity, which is better compared to the 88.4% accuracy and 40% specificity is the current system.

In May 2015, Md. Badrul Alam Miah and Mohammad Abu Yousuf proposed a Neural Network based CAD system for early detection and diagnosis. The existing systems using ANN and fuzzy clustering, IP, Curvelet transform, multinomial Bayesian algorithm, back propagation, gray-coefficient mass estimation and SVM are the basis of these observations. The goal is to create a fast and robust, more accurate system having a rotation, scaling and translation variant feature extraction [3]. A dataset of 300 images acquired from hospitals is used. Steps in proposed system are Image acquisition, processing, binarization, segmentation using thresholding, feature extraction and neural network classification. Steps in Image processing are grayscale conversion, normalization, noise reduction, binarization and removing unwanted portion of image. Feature extraction uses features like center of image, ratio of height to width, average distance between black pixels and the center, etc. Neural networks is used for classification with two outputs. The system gives an accuracy of 96.67%, higher than all existing systems.

In March 2014, Prof. Sanjeev N. Jain and Bhagyashri G. Patil proposed few methods to detect tumorous cells from CT Scans of Lungs. The purpose of this paper is to find the tumorous cells and give more accurate result by using different segmentation techniques such as thresholding and watershed segmentation. In thresholding, a threshold value is set to differentiate region of interest from the background. If the pixel value is greater than the threshold value then it belongs to the object else it is in the background. Thus the region of interest can be extracted by using thresholding approach. In watershed segmentation, the background and the object are separated using different markers : internal markers associated with object of interest and external markers associated with background. It is simple, intuitive and fast method [4]. According to the research, watershed segmentation approach has more accuracy (85.27%) than thresholding approach (81.24%).

In 2017, authors Pooja R. Katre and Anuradha Thakare described the various image processing techniques for detecting lung cancer. In their proposed approach for noise removal and enhancement, method Median Filter is used. The best part about median filter is it removes noise without blurring the image. It preserves the edges of the regions [5]. Other techniques remove

the noise and give blur image as output. It is used to remove salt and pepper noise from the image. Medfilt2 function of MATLAB is used for this technique. In enhancement stage gabor filter is used as it gives improved result compared to fast fourier and auto enhancement. In this paper, several image processing techniques are described which help to process the image and extract the features of image and then the image is classified. The purpose of this paper is to detect the tumour at an early stage. Initially, CT scan images are taken as input from the database. Later, the second stage is Image pre-processing in which smoothing, image enhancement and image segmentation takes place. Noise in the image is removed in this stage. Median filter gives the best result as it removes noise. Next is the feature extraction stage in which the area, perimeter, Eccentricity of the image is calculated. Support Vector Machine algorithm is used to classify the data. The features mentioned above help to identify the size of the tumour and from that, the stage of the cancer is detected.

In 2017, from China, Lei Fan and his group of researchers made use of deep learning algorithm for CAD lung cancer detection, realizing the importance of CAD for diagnosing the increasing number of cancer related deaths in China. In this paper, image processing is not applied on CT scans of lungs. The images are directly fed as an input to convolutional neural network which consist of two convolutional layers, two max pooling layers, one fully connected layer and one output layer [6]. Rectified linear unit (ReLU) are applied between convolutional and max pooling layers. The system gives an overall accuracy of 67.7%. It concludes that Support vector machines has lower classification accuracy than 3D convolutional neural network in the same number of input samples.

In 2017, authors Qing Wu and Wenbing Zhao published a paper to propose a CAD based Lung cancer detection system. In their system, neural network based algorithm, EDM (Entropy Degradation Method) is proposed to detect SCLC (small cell lung cancer) from CT scan images [7]. The training and testing data are lung CT scan images which are provided by the National Cancer Institute. Five scans from each group are randomly selected to train the model. The images with SCLC are labelled as cluster 1 while others are labelled as cluster 0. The paper proposes early stage detection of lung cancer.

In another paper by authors Anita Chaudhary and Sonit Sukhraj Singh, the proposed CAD system has techniques to detect the tumour at an early stage. The paper consists of three important processes which are used throughout. The use of MATLAB is done at every phase of the project. Several image processing techniques such as, image pre-processing, segmentation and feature extraction have been used. In image enhancement stage, Gabor filter enhancement technique, Auto Enhancement Technique and Fast Fourier Transform Technique is used out of which the Gabor Filter Enhancement Technique is the most suitable one. According to the research, the Watershed segmentation approach is more accurate than the Thresholding segmentation approach [8]

In 2017, Deep Prakash Kaucha and others proposed a system to increase accuracy, sensitivity and specificity in LCD using CAD system. This study aims at maximizing effectiveness from strategic deployment of different techniques at different stages [9]. The system studies the issues faced in Image Processing and related works on early detection of Lung Cancer. The existing system with its flow siagram is studied, which uses K Means for clustering, Gray Level Co-occurrence Matrix (GLCM) for feature extraction and Back Propagation for classification. The proposed model uses Discrete Wavelet Transform for segmentation, GLCM for feature extraction, SVM for classification with database from LIDC and implementation in MATLAB. ROC curve for True Positive and False Positive in existing and proposed systems show an improvement in performance. The system shows an accuracy of 95.16%, sensitivity of 98.21% and specificity of 78.69%.

In 2017, authors S. Kalaivani, Pramit Chatterjee, Shikhar Juyal and Rishi Gupta of VIT University, Tamil Nadu proposed a simple CAD system for Lung Cancer Detection, knowing the importance of research in the field of Cancer research. The paper talks about Lung Cancer and its effects. The data used is from the The Cancer Imaging Archive. The images are converted to grayscale, improved using Histogram Equalization and segmented using Thresholding. Feature Extraction is performed by calculating certain properties in the binary image [10]. This is done using the regionprops function in MATLAB. Back Propagation network is used for classification as ANN can perform multiple parallel calculations. This simple system gives an accuracy of 78% and can be helpful to radiologists.

In 2014, Prof. L.M. Deshpande and Mr. Vijay A. Gajdhane acquired CT Scan images from various hospitals for lung cancer detection using Image Processing. Noise reduction filters, thresholding, segmentation using watershed segmentation and feature extraction were studied to propose a methodology using MATLAB. Input images were passed through smoothing, enhancement and then segmented using marker-controlled watershed segmentation [11]. Features like area, perimeter, eccentricity were extracted for the dataset and classification was done using Maximum margin classifier.

In 2016, Prajwal Rao, Nishal Pereira and Raghuram Srinivasan proposed a Convolution Neural Network based system for the screening and detection of Lung Cancer from CT Scan Images. Data from the LIDC was used for the system. The data and its label is the input for the system. The CNN architecture consists of one input layer, two convolution layers, a max pooling layer and a fully connected layer and an output layer [12]. In the first convolution layer, the output is 78 features. After the second convolution layer, the CNN network learns the features hierarchically and avoids the need for a specific feature engineering. The max pooling layer reduces the data dimensions. The fully connected layer gives two outputs by combining all the features for a input image. The system was compared to an ANN system with 50 neurons for accuracy. An accuracy of 75% is obtained in comparison the ANN based system.

IV. ANALYSIS

Table 1 gives the analysis of techniques and methods used in literature survey on Lung Cancer detection and Image Processing on CT Scan Images.

**TABLE 1
ANALYSIS TABLE**

Sr. No.	Title	Advantages	Open Challenges	Accuracy (%)
1	Lung cancer detection using Image Processing technique [1].	Compares different methods and algorithms in each Image Processing step. Gives results/comparisons precisely.	Complexity / resources required are not distinctly mentioned. No background on further classification.	-
2	Lung Cancer Detection using CT scan images [2].	Proper calculation of results. Well-defined feature extraction.	SVM is used. Newer, better approaches are now available using deep learning.	88.4
3	Detection of lung cancer from CT Image using Image Processing and Neural Networks [3].	Better classifier than standard binary classifiers. Rotation and scaling variant feature extraction.	Image Processing steps involved, source of input data not mentioned.	96.67
4	Cancer cells detection using Digital Image Processing Methods [4].	Watershed segmentation is better than thresholding.	Complexity and processing time is more. Maintains two different markers for region of interest and background.	-
5	Detection of Lung Cancer Stages using Image Processing and Data Classification Techniques [5].	Median Filter preserves the edges of the region of interest.	Median filter is difficult to apply on CT scans as the images are in grayscale.	-
6	Lung nodule detection based on 3D	3D CNN give better classification results than SVM for same number of input	CNN is time consuming and requires more computing power	67.7

	Convolutional Neural Networks [6].	samples.	and higher graphics.	
7	Small- Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm [7].	Focus on small cell lung cancer detection. Accurate dataset. Output in clusters of cancerous/non-cancerous.	Stages / techniques used not clearly explained and results are inaccurate.	-
8	Lung Cancer Detection on CT images by using Image Processing [8].	Image processing techniques mentioned with comparisons and proper flow of stages in Image Processing.	Classification not mentioned..	-
9	Early Detection of Lung Cancer using SVM Classifier in Biomedical Image Processing [9].	Comparisons in accuracy are complete and proved. The system uses implementations which are sure to improve the performance.	SVM is used instead of deep learning techniques. Image Processing can be improved by using different set of techniques.	95.16
10	Lung Cancer Detection Using Digital Image Processing and Artificial Neural Networks [10].	Simple CAD system using Back Propagation. Good efficiency because of its simple system flow diagram using easier techniques and algorithms.	The accuracy of the system is not very high, and for a CAD for a lung cancer detection, accuracy is more important than efficiency.	78
11	Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques [11].	The changes in input after applying the Image Processing technique is shown after each stage. Cancer stage criteria in terms of features extracted is mentioned.	The proposed system does not talk about the accuracy and efficiency. SVM which is a better classifier can be implemented.	-
12	Convolutional Neural Networks for Lung Cancer Screening in Computed Tomography (CT) Scans [12].	Images from the LIDC dataset is used. The architecture of the CNN network is given in the paper. Comparisons between results of ANN, linear network and the proposed system is shown in the form of graphs.	The accuracy of the system can be improved further. This system proposes first steps towards automation of feature extraction using CNN and further improvements can be made.	75

Based on the analysis of various research papers, deep learning, which is the latest development in the field of Artificial Intelligence, is deemed optimal for computer aided detection systems. Accuracy using CNN in the present implemented systems is around 67%. This can be improved by researchers by using a well maintained database for training models, better defined layers in CNN for better feature extraction and using preprocessing measures on data for a better result.

V. CONCLUSION

The above papers discussed deep learning techniques; machine learning techniques like SVM, ANN, fuzzy clustering. Different techniques gave different results on different datasets giving varied accuracy. A well maintained LIDC dataset was used.

Deep Learning is a branch of Machine Learning that deals with the use of deep neural networks for decision making and classification. There are many systems that perform detection of cancerous tumor in CT Scan images. These systems make use of various classification algorithms. But existing systems using deep learning algorithms have low accuracy. Accuracy is very important in detection of diseases like cancer.

Deep Learning is more powerful than conventional Machine Learning algorithms due to the computations in hidden layers. This Deep Learning based CAD has promising applications in the field of Biomedical Technology.

ACKNOWLEDGEMENTS

We would like to express a deep sense of gratitude towards our guide Prof. Janhavi Sangoi and Co-Guide Prof. Vinit Raut, Computer Engineering Department for their constant encouragement and valuable suggestions. The work that we are able to present is possible because of their timely guidance and support.

REFERENCES

- [1] Md. Badrul Alam Miah et. al, "Detection of lung cancer from CT Image using Image Processing and Neural Networks", 2015 International Conference on Electrical Engineering and Information Communication Technology, (ICEEICT), Dhaka, 2015, pg. 1-6.
- [2] P. R. Katre and A. Thakare, "Detection of Lung Cancer Stages using Image Processing and Data Classification Techniques", 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017, pg. 402-404.
- [3] L. Fan et. al, "Lung nodule detection based on 3D Convolutional Neural Networks", 2017 International Conference on the Frontiers and Advances in Data Science (FADS), pg.7-10.
- [4] Mr. Vijay A. Gajdhane and Prof. L. M . Deshpande, "Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16 issue 5 (Sept – Oct 2014), pg. 28-35, e-ISSN 2278-0661, p-ISSN 2278-8727.
- [5] Mokhled S. Al Tarawneh, "Lung Cancer Detection Using Image Processing Techniques", Leonardo Electronic Journal of Practices and Technologies, Issue 20, pg. 147-158, January June 2012, ISSN 1583-1078.
- [6] Suren Makaju et. al, "Lung Cancer Detection using CT scan images.", 6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8th December 2017, Kurukshetra, India.
- [7] Sanjeev Jain and Bhagyashri G. Patil, "Cancer cells detection using Digital Image Processing Methods", International Journal of Latest Research in Science and Technology, Volume 3(4), March 2014, pg. 45-49.
- [8] Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm", 2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC), Budapest 2017, pg. 88-91.
- [9] D. P. Kaucha et. al, "Early Detection of Lung Cancer using SVM Classifier in Biomedical Image Processing", IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017), Chennai, 2017, pg. 3143-3148.
- [10] Chaudhary and S. S. Singh, "Lung Cancer Detection on CT images by using Image Processing", 2012 International Conference on Computing Science, Phagwara, 2012, pg. 142-146.
- [11] S. Kalaivani et. al, "Lung Cancer Detection Using Digital Image Processing and Artificial Neural Networks", International Conference on Electronic, Communication and Aerospace Technology (ICECA) Coimbatore, 2017, pg. 100-103.
- [12] P. Rao et. al, "Convolution neural networks for lung cancer screening in computed tomography (CT) scans", 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, 2016, pg. 489-493.
- [13] https://en.wikipedia.org/wiki/Median_filter , last accessed on 12/10/2018.
- [14] <https://in.mathworks.com/help/images/marker-controlled-watershed-segmentation.html>, last accessed on 12/10/2018.
- [15] https://www.google.co.in/search?q=convolution+neural+networks&oq=convolution+neural+networks&gs_l=psy , last accessed on 12/10/2018.
- [16] Konrad Paul et. al, "Rethinking the role of Machine Learning in Biomedical Science", Frontiers of Engineering: Reports on Leading-Edge Engineering from 2017 Symposium, Washington DC, January 2018.