Techniques for Lung Cancer Detection from CT Image

Sugandha Saxena, S. N. Prasad, Bhavanishankar

Abstract: The most lethal disease found in the medical field is lung cancer and early detection of this disease has become a challenge for many doctors and diagnostics. The lung cancer contributes over 15.3% of the total number of new cases diagnosed in the recent years. Smoking and pollution are considered as the major causes of lung cancer. At present, there are huge number of tests available to detect lung cancer such as PET Scan, Computerized Tomography (CT) Scan and X-ray etc. are used to diagnose the disease. By x-ray the picture of the lungs may uncover the unusual mass or nodule. A further developed adaption found in CT scan which can uncover the small lesions in the lung that probably won't be distinguished with X-ray. Biopsy tests are done for detailed diagnosis of the disease. For accurate and better results, a data mining techniques, machine learning algorithms or deep learning algorithms could be used in the laboratories. In this survey, we have elaborated various existing techniques used so far.

Keywords: Lung cancer, data analytics, machine learning algorithm, deep learning algorithms.

I. INTRODUCTION

In present scenario, the risk of lung cancer is increased resulting a huge threat to the generation. The main sufferers are those who smoke, exposed to polluted air, radiation etc. The estimated casualties due to the cancer disease are above 15 lakh and over 20% because of lung cancer [1]. Doctors can identify the maligned part by using the different techniques available. There has been a tremendous improvement in performance, speed and other parameters observed because of a significant research contribution by the scientists, researchers in the area of Convolution Neural Network (CNN), Artificial Neural Network (ANN) using certain algorithms. From convolutions into fully connected layers [16], to the addition of dropout layers [11], to improvement [14, 2], the way to deal with profound learning is always showing signs of change and improving.

John Holland introduced the Genetic algorithms (GA) in the early 1970s and these methods will try to mimic the process of natural selection to discover the answer for certain streamlining and research issues.

Convolution Neural Networks were elaborated in 1989 with LeNet-5 [16].The design of LeNet-5 was to handle a variance among identically labeled data by using "local receptive fields, shared weights, and spatial sub-sampling."

Sugandha Saxena, School of ECE, REVA University, Bangalore, India. Email: sugundha.saxena@reva.edu.in

S. N. Prasad, School of ECE, REVA University, Bangalore, India. Email: prasadsn@reva.edu.in

Bhavanishankar, Department of Computer Science and Engineering, RNSIT, Bangalore, India. Email: bsharsh@gmail.com

Even though it is found to be more effective, the change has been observed with AlexNet [15] which elaborated the idea of using CNNs in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC).

GoogLeNet[24] presented the inception module which was planned considering two thoughts. Those are the estimation of a sparse structure with spatially repeated thick parts and to hold the computational intricacy under tight restraints. Extracting features from the separate cluster of information by convolution filters is the biggest benefit of inception architecture. When one repeats them spatially, they estimate the optimal sparse structure with thick components. Inception architectures bearing the principle of visual information processing at various stages in large scale. The conclusion is that the inception module would be having three times faster performing networks compared to non-inception architectures.

II. RELATED WORK

Recently, Google Brain contributed towards the usage of repeated neural networks to look for ideal designs using reinforcement learning [25]. In a project carried out at the University of Bern, a deep CNN design has been created which classifies the lung nodule with an accuracy of 86%. In this project, a dataset has been used which contained several lung slices from CT images [4]. Six different lung cancer variants images were taken and experiments were carried out. Their dataset consisted of around 2100 different diseases. To deal with such a varied dataset, tree taxonomy has been developed by the researchers. To do the extreme classification, dynamic tree taxonomy and its associated are generated based on the number of examples rather than final diagnoses. This resulted in have around 800 classes, instead of 2100.

In one of the project at the Federal University of Parana, a classification on CNN has been made by considering the images of cell slides of breast cancer patient [22]. They used AlexNet to distinguish microscopic biopsy images of breast tumors as malignant or benign by using the BreaKHis dataset [21]. It comprises of 4 images each with various degrees of magnification. Few techniques have been introduced to handle the high resolution nature.

In [7], researchers have achieved human specialist performance in several tasks. Accurately identifying lung cancer in the early stages is preferred; otherwise it may lead to a disaster. CT scan is done by radiologists to identify lung cancer using CAD (Computer Aided Diagnosis).

Earlier, detection of lung cancer was very challenging and radiologist used to take enormous amount of time, if they manually detect the nodules by using CT scans. Therefore, an automatic system is required which help the radiologist to take the decision fast while reducing their work load. As the technology advanced, this era has received many machine learning approaches to resolve this problem namely artificial neural network, support vector machine etc [8, 9, 10, 11, 12, 13]. Studies showed that very little work has been done using Convolution Neural Network (CNN) approaches to classify nodules in CT images. This is mainly due to the limited size of labeled databases.

III. DIFFERENT METHODS USED FOR LUNG CANCER DETECTION

By using different approaches, detection of lung cancer has become easy. These new systems will assist the radiologist and help in detecting and classifying the disease with high precision. Image processing, Machine learning and deep learning are different methods available to detect lung cancer. Table-I shows comparison between different approaches for lung cancer detection.

A. Image Processing Methods

In this method detecting lung cancer involves four basic steps-Image dataset, image enhancement, and image segmentation and feature extraction. The first stage involves collection of dataset. The dataset should contain CT scan images. Usually for lung cancer detection, CT scan images are preferred. CT scan images are high resolution and very clear with low noise distortion. Second stage is preprocessing which helps in enhancing the image quality. Many methods such as image blurring and filters such as wiener filters are used for denoising the image. Third stage will put in various image segmentation algorithms to segregate the image into large number of pieces to retrieve the necessary information. The main advantage of segmentation is to obtain better orientation in the image. In segmentation, a CT scan image is divided into number of pieces. In this way, analyzing the image becomes very easy and chances of human error while detecting the lung cancer reduces. Using these methods, many researchers have presented much efficient lung cancer detection. Last step is feature extraction which helps in extracting only the required features from the image. Various algorithms can be used for feature extraction such as binarization approach and masking approach.

B. Machine Learning Methods

In cancer detection machine learning approach has been used extensively. It is a sphere of knowledge where the computer learns from the instances to perform various operations. Various learning algorithms are applied on training data set such as artificial neural networks, SVM, Logistic Regression to produce an output. Machine learning models are adaptable and produce precise output by using an optimized algorithm. Recently a model was proposed for nodule detection by using CT scan images while applying watershed segmentation [27]. Gaussian filter and median filter was used for image processing and support vector machine which helps in identifying the nodule as threatening or not.

C. Deep Learning Methods

Though machine learning approach has proven to be an effective method but the processing time of raw images is to high. Usually medical related applications have large dataset which can lead to slow tuning of the images. Therefore, this limitation of machine learning is now taken care by deep learning methods. Deep learning is very promising in the field of medical applications including radiology, dermatology and pathology. Deep learning algorithms automatically identify the complex pattern in input data and then produce an output in a significant manner. These algorithms are fast and they try to learn multiple levels of abstraction, representation, feature extraction and information from the set of input medical images. In machine learning, multi resolution methods such as Wavelet Transform have been greatly adopted for feature extraction. This method is very advantageous over other feature extraction methods.

Recently, a Convolution Neural Network (CNN) based model was created to identify a lung nodule as cancerous or not [29]. Median filter was used in the preprocessing stage to obtain high clarity image by minimizing the adverse effects of degradation during learning [29]. There was another model developed by the researchers for lung nodule detection which used ensembles of CNN and multiple preprocessing methods in order to increase the accuracy. A voting system is used which produce the output when the two CNN are agreed else the output is negative[28].In one of the paper, the author has utilized the versatile adaptive distance based threshold algorithm for segmentation and Fisher Linear Discriminant Classifier to distinguished the computed features. The efficiency of this system detected was 78.1%.

Table- I: Comparison between different approaches for lung cancer detection.

S.No	Sudy	Method	Result
1.	N.Singh and A. Asunthana[26]	Canny Edge Detection and Gabor Filter for Image preprocessing. Super Pixel Segmentation	Fast performance and suitable for real time detection.Reduces image complexity and gives higher output in comparison to other algorithms.
2.	Suren Makaju, P.W.C. Prasad, Abeer Alsadoon, A. K. Singh, A. Elchouemi[27]	Watershed Segmentation Dataset provided by LDIC	Achieved accuracy of 88.4%
3.	Suren Makaju, P.W.C. Prasad, Abeer Alsadoon, A. K. Singh, A. Elchouemi[27]	Watershed Segmentation and SVM for classification.Dataset provided by LDIC	Helps in proper segmentation of cancer nodules with achieved accuracy of 92%

| 4. | Allison M Rossetto and Wenjin Zhou[28] | Used ensembles of Convolution Neural Network(CNN) with preprocessing pipeline method.Database provided by the Kaggle Data Science Bowl | High accuracy of 97.5% and low percentage of false reduction (<10%) |

IV. DATA ACQUISITION AND PREPROCESSING

The main challenge to develop the intelligent models for detecting lung cancer is collection of dataset images for trained model. Several image based scans are available which help to detect the lung cancer. CT images are preferred because they can give high resolution for the nodules and tumors. Also, they have the ability to form 3D images of the lungs. MRI also helps in detecting the location of the tumor. But MRI is not suitable for taking pictures of moving body parts like lungs as it moves with each breath.

There is a large public data available for lung CT images. Researchers can obtain the images from: Lung Image Database Consortium (LIDC) or Medical Image Database or Early Lung Cancer Action Program (ELCAP).As per the survey, it is observed that LIDC is considered to be the standard benchmarking dataset. LIDC is a repository of CT scan images prepared by NCI (National Cancer Institute) for researchers so that new biomedical devices can be designed. Another advantage is that all LIDC images are stored as slices in the database and are available in DICOM format.

Preprocessing of an image is a very important step as it helps in magnifying the clarity and quality of an image by reducing the unwanted noise. CT scan images not only contain lung but other substances surrounding it such as air, blood, water and bones. In CT scan images, other than lungs, presence of other substance is not important so we need to exclude them in order to achieve the high accuracy. Types of filters can be used in this stage such as Median filter, Gabor filter, selective enhancement filter and so on. Most of the researchers will use Median filter [34][35] in the preprocessing stage as it does not distort the edges.

V. SEGMENTATION

This is the second phase of the process which is very challenging. Segmentation is a process which split the image into several parts and helps to detect the nodule with high accuracy. This step is significant as there are many befuddling areas outside the lung area. So this process helps to enhance the precision and detect the cancer/tumor effectively at early stages. There are a variety of methods which will do segmentation such as Adaptive threshold method, active contour model, Thresholding method, Otsu method, Gaussian mixture model etc. Table II & III shows the various methods used and the performance obtained for lung cancer detection.

A. Thresholding Method

This is a popular method which is used in segmentation. This method helps in segregating the lung region from the CT scan image. Threshold value is not fixed and it will vary across the different images. This variation in threshold value may be because all the images background is not grey circular region. The accuracy of this method depends completely on acquisition type and its protocol. The threshold value is chosen between lung pixel value and dense tissue pixel value. So in this method, based on threshold value gray scale image is turned into binary image [35].

B. Gaussian Mixture Model

This method is used to model the dispersion of intensity value of the voxel of the image so that internal structure and pulmonary parenchyma can be differentiated.GMM is used because these are well acquainted statistical inference technique and this model is strong enough in choosing the component distribution and hence produces a density estimation for each cluster.

Table- II: Comparative Analysis of various stages included in each selected work

Study	Preprocessing	Lung segmentation	Feature Extraction
Wook-Jin Choi, Tae-Sun Choi[30]	No	Large volume segmentation	Angular Histograms approach
Alex Martins Santos et al.[31]	No	Gaussian Mixture model	Shannon Entropy
Jinsa Kuruvilla, K. Gunavathi [32]	No	Otsu Thresholding	-
D. Kumar, A. Wong, and D. A. Clausi[33]	No	-	Autoencoder
Sarah Soltaninej et al.[34]	Median Filter	Active contour modeling	2D stochastic feature
Md. Badrul Alam Miah et al[35]	Median filter	Thresholding Method	-

C. Active Contour Model

This method is used in lung segmentation. Every initial stage of CT image is applied with masking technique. For every next slice, the lung contour of each image will become the initial mask[34].

D. Otsu Thresholding

Using morphological operations, the lung is separated from the CT image. Firstly the grayscale image is changed into binary image. Then according to threshold value all the pixels in the input image having intensity value more than threshold value are replaced by 1 and if pixel having intensity is less than threshold value is replaced by 0. Threshold value can be calculated using Otsu method because it is best in maximizing the between -class measure which is an important parameter in statistical Discriminant analysis.

Table- III: Performance parameters in each selected work

Study	Sensitivity	Specificity	Accuracy
Wook-Jin Choi, Tae-Sun Choi[30]	97.5%	-	-
Alex Martins Santos et al.[31]	90.6%	85%	88.4%
Jinsa Kuruvilla, K. Gunavathi [32]	91.4%	93.3%	100%
D. Kumar, A. Wong, and D. A. Clausi[33]	83.35%	-	75.01%
Md. Badrul Alam Miah et al[35]	-	-	96.67%

Table- III: False Positive reduction in each selected work

Study	Dataset	Classifier	False Positive Reduction
Wook-Jin Choi, Tae-Sun Choi[30]	LIDC	SVM	6.76 false positive per scan
Alex Martins Santos et al.[31]	LIDC	SVM	1.17 false positive per scan
Jinsa Kuruvilla, K. Gunavathi [32]	LIDC	Feed Forward with Back propagation Neural Network	No
D. Kumar, A. Wong, and D. A. Clausi[33]	LIDC	CAD Classifier system	0.39/patient over a 10 fold cross validation
Sarah Soltaninej et al.[34]	-	KNN	5.63FP/scan
Md. Badrul Alam Miah et al[35]	-	Multilayer Feed Forward Network	No

VI. FEATURE EXTRACTION

After segmentation, feature extraction is an important step in CAD systems. This step helps us to extract the features of the nodule which in turn supports more understanding of the image. Various features such as area, shape, geometry, perimeter, eccentricity, and diameter are calculated. Later these features are utilized as training features to build a classifier. As shown below many approaches and algorithms can be used for feature extraction.

A. Angular Histogram Approach

Many researchers have used histogram approach to deal with the shape of the object. The shape of the target component can be anticipated dependent on the direction likelihood of surface normal. Hessian grid is utilized on each voxel of image so as to compute the surface ordinary.

B. Shannon Entropy

In this method, texture features such as Shannon entropy is used to detect nodule from non-nodule. Entropy of any discrete source is determined from the probability distribution. The probability of distribution is equivalent to the probability of the system in certain state.

The Shannon entropy S can be described as follows:

$$S = -\sum_{i=1}^{k} p_i \ln p_i$$

C. Autoencoder

This type of encoder is used for feature extraction purpose. Autoencoder is a type of two layer network which encode the input data into the latent space. To perform this operation it makes use of liner or non linear transformations. Type of activation function used in autoencoder is sigmoid or hyperbolic tangent functions.

VII. RESULTS AND DISCUSSIONS

In this paper we are analyzing the different procedures for preprocessing, segmentation, feature extraction and classification using different machine learning algorithms. For classification of detected nodules will be carried out using Maximum Sensitivity (MS) algorithm to find out whether the nodule will be malignant or not. The flow chart for methodology of lung cancer detection system is given as follows:
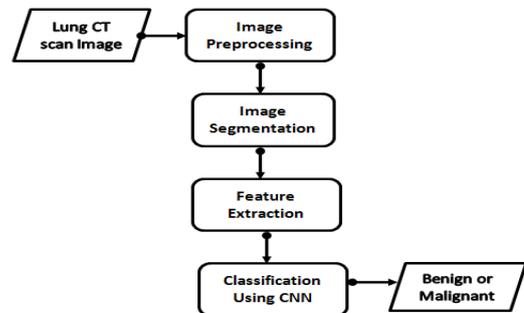
Fig.1 Flow Chart for Lung Cancer Detection

VIII. CONCLUSION

In this paper, comparative analysis has been done for different techniques which are used for lung cancer detection. Therefore, we proposed convolution neural network based system as the accuracy achieved through CNN is very high. Different classifier techniques can be clubbed with different segmentation and feature extraction methods to obtain the accuracy for detecting the lung nodule. During survey, it has been observed that only CT scan images are appropriate to obtain accurate results.

Therefore, modeling such devices with high precision is required so as to distinguish the lung malignancy at the beginning stage.

REFERENCES

1. G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
4. B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
5. E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," IEEE Trans. Antennas Propagat., to be published.
6. J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," IEEE J. Quantum Electron., submitted for publication.
7. Andre Esteva et al. "Dermatologist-Level Classification of Skin Cancer With Deep Neural Networks". In: Nature (2017), pp. 115–118.
8. L. G. Hafemann, L. S. Oliveira, and P. Cavalin. "Forest Species Recognition Using Deep Convolution Neural Networks". In: 2014 22nd International Conference on Pattern Recognition. Sept. 2014, pp. 1103–1107. doi: 10.1109/ICPR.2014.199.
9. Kaiming He et al. "Deep residual learning for image recognition". In: Computer Vision and Pattern Recognition, 2016.
10. Seth Hendrickson. MarI-O. https://github.com/pakoito/MarI-O. 2015.
11. G. E. Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: arXiv preprint arXiv:1207.0580 (2012).
12. Sumati jaggi, Ravi Bhushan, and Davender Malhotra. "Guidelines to Decide the Encoding Scheme Used For G.A." In: International Journal of Advanced Research in Computer Science and Software Engineering Vol. 3, Issue 8 (2013), pp. 1436–1440.
13. Yangqing Jia et al. "Caffe: Convolutional Architecture for Fast Feature Embedding". In: arXiv preprint arXiv:1408.5093 (2014).
14. Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: CoRR abs/1412.6980 (2014).
15. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks". In: Neural Information Processing Systems (2012).
16. Yann LeCun et al. "GradientBased Learning Applied to Document Recognition". In: Pro- ceedings of the IEEE Vol. 86, Issue 11 (1998), pp. 2278–2324.
17. Risto Miikkulainen et al. "Evolving Deep Neural Networks". In: CoRR abs/1703.00548 (2017).
18. Joao Carlos Figueira Pujol and Riccardo Poli,"Evolving the Topology and the Weights of Neural Networks Using a Dual Representation. "vol 8,issue 1,pg 73-84,1998.
19. Wolfram Schiffmann, Merten Joost, and Randolf Werner. "Performance Evaluation of Evo- lutionary Created Neural Network Topologies". In: Springer-Verlag London Vol. 2 (1990), pp. 274–283.
20. Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: CoRR abs/1409.1556 (2014). url: http://arxiv.org/abs/1409. 1556.
21. F. A. Spanhol et al. "A Dataset for Breast Cancer Histopathological Image Classification". In: IEEE Transactions on Biomedical Engineering 63.7, pp. 1455–1462,2016.
22. F. A. Spanhol et al. "Breast cancer histopathological image classification using Convolutional Neural Networks". In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2560–2567,2016.
23. Kenneth O. Stanley and Risto Miikkulainen. "Evolving neural networks through augmenting topologies". In: Journal of Evolutionary Computation Vol. 10 Issue 2 (2002), pp. 99–127.
24. Christian Szegedy et al. "Going deeper with convolutions". In: Computer Vision and Pattern,Recognition (2015).
25. Barret Zoph and Quoc V. Le. "Neural Architecture Search with Reinforcement Learning". In: CoRR abs/1611.01578 (2016).
26. N. Singh and A. Asuntha, "Research Article Image Processing used for Lung Cancer Detection in Medical Imaging," vol. 8, no. 4, pp. 1044–1049, 2016.
27. S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Lung Cancer Detection using CT Scan Images," Procedia Comput. Sci., vol. 125, no. 2009, pp. 107–114, 2018.
28. A. M. Rossetto and W. Zhou, "Deep Learning for Categorization of Lung Cancer CT Images," Proc.-2017 IEEE 2nd Int. Conf. Connect. Heal. Appl. Syst. Eng. Technol. CHASE 2017, pp. 272–273, 2017.
29. S. Sasikala, M. Bharathi, B. R. Sowmiya,"Lung Cancer Detection and Classification Using Deep CNN",International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5S4, February 2019.
30. Wook-Jin Choi, Tae-Sun Choi,"Automated pulmonary nodule detection based on three-dimensional shape-based feature descriptor", Computer methods and programs in biomedicine,pg 37-54,2014.
31. Alex Martins Santos a et al.,"Automatic detection of small lung nodules in 3D CT data using Gaussian mixture models, Tsallis entropy and SVM",Engineering Applications of Artificial Intelligence,pg 27-39,2014.
32. Kuruvilla, K. Gunavathi,"Lung cancer classification using neural networks for CT images",Computer methods and programs in biomedicine,pg 37-54,2014.
33. Devinder Kumar,Alexander Wong,David A. Clausi,"Lung Nodule Classification Using Deep Features in CT Images",12th Conference on Computer Robot Vision,pg 133-138,2015.
34. Sarah Soltaninejad et al. "Lung Nodule Detection by KNN classifier and Active Contour Modelling and 3D Vizulization",The 16th CSI Internatioal Symposium on Artificial Intelligence and Signal Processing, 2012.
35. Md. Badrul Alam Miah, Mohammad Abu Yousuf," Detection of Lung Cancer from CT Image Using Image Processing and Neural Network", 2nd Int'l Conf on Electrical Engineering and Information & Communication Technology (ICEEICT) ,2015.

AUTHORS PROFILE

Sugandha Saxena is Assistant Professor of Electronics and Communication Engineering at REVA University with an experience of 6years in teaching. She completed M.Tech in VLSI design from Amity University, Noida. She accomplished M.Tech thesis from CSIR, Pillani. Currently, she is pursuing Ph.D from REVA University. Her area of interest is CMOS VLSI Design, Semiconductor Devices and Artificial Intelligence.

Dr. S. N. Prasad is Professor of Electronics and Communication Engineering & Special Officer (Evaluation) at REVA University, Bangalore. He has completed B.E in Electronics and Communication Engineering from Mangalore University and M.Tech in Power Electronics from Visvesvaraya technological University. He has more than 20 years of teaching and 3 years of industry experience. He has many publications in reputed national/international journals and conferences which are having very good impact factors. His research interests include Artificial Intelligence, Embedded Systems, Real Time Systems, Signal Processing, Power Electronics and Automotive Electronics. He is a member of ISTE(MISTE) India, member IE (MIE)India and member of IETE

Dr. Bhavanishankar K. is an Assistant Professor of Computer Science and Engineering department at RNSIT, with an experience of 16 years in teaching and 7 years in research (inclusive). He has published 8 papers in international journals. His area of interest is image processing, algorithms, database management and artificial intelligence.