International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT20631110

A Survey on Multistage lung cancer Detection and Classification

Jay Jawarkar[1], Nishit Solanki[2], Meet Vaishnav[3], Harsh Vichare[4], Dr. Sheshang Degadwala[5]
[1,2,3,4]U.G. Scholar, Sigma Institute of Engineering, Vadodara, Gujarat, India
[5]Associate Professor, Sigma Institute of Engineering, Vadodara, Gujarat, India

ABSTRACT

Article Info
Volume 6, Issue 5
Page Number: 313-320
Publication Issue :
September-October-2020

Earlier, Lung cancer is the primary cause of cancer deaths worldwide among both men and women, with more than 1 million deaths annually. Lung Cancer have been widest difficulty faced by humans over recent couple of decades. When a person has lung cancer, they have abnormal cells that cluster together to form a tumor. A cancerous tumor is a group of cancer cells that can grow into and destroy nearby tissue. It can also spread to other parts of the body. There are two main types of lung cancer:1. Non-small cell lung cancer, 2. Small cell lung cancer. Non- small cell lung cancer has four main stages. In this research we are classifying four stages of lung cancer. Lung cancer detection at early stage has become very important. Currently many techniques are used based on image processing and deep learning techniques for lung cancer classification. For that lung patient Computer Tomography (CT) scan images are used to detect and lung nodules and classify lung cancer stage of that nodules. In this re- search we compare different Machine learning (SVM, KNN, RF etc.) techniques with deep learning (CNN, CDNN) techniques using different parameters accuracy, precision and recall. In this Research paper we com- pare all existing approach and find our better result for future application.

Article History
Accepted : 10 Oct 2020
Published : 30 Oct 2020

Keywords: SVM, KNN, CNN, RNN, Lung Cancer, Stages, CT Image, Nodules, Deep Learning

I. INTRODUCTION

There are two major types of lung cancer, non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). Staging lung cancer is based on whether the cancer is local or has spread from the lungs to the lymph nodes or other organs.

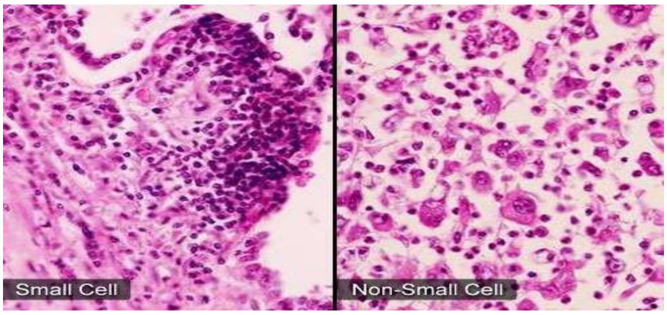

Figure 1: Lung Cells

Copyright: © the author(s), publisher and licensee Technoscience Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited

Because the lungs are large, tumors can grow in them for a long time before they are found. Even when symptoms—such as coughing and fatigue—do occur, people think they are due to other causes. For this reason, early-stage lung cancer (stages I and II) is difficult to detect.

Earlier, Lung cancer is the primary cause of cancer deaths worldwide among both men and women, with more than 1 million deaths annually. Lung Cancer have been widest difficulty faced by humans over recent couple of decades. When a person has lung cancer, they have abnormal cells that cluster together to form a tumor. A cancerous tumor is a group of cancer cells that can grow into and destroy nearby tissue. It can also spread to other parts of the body.

In this research paper, we summarize different types of lung cancer stages for classification. For that different feature and deep learning approaches introduce new strategy combine shape and texture feature. For shape feature we will use segmentation based on clustering and after that deep classification for future stage prediction.

II. RELATED WORKS

In [1] Janee Alam1, Sabrina Alam2, Alamgir Hossan3. They have described about lung cancer multistage detection using these methods are SVM Classifier, GLCM feature and Water shaded Transform. This paper limitations are, small dataset, low accuracy and used large feature dimensions.

In [2] Emine CENGİL, Ahmet ÇINAR. They have described about deep learning approach to lung cancer and they used Convolution Neural Network (CNN). This paper has some limitations are classify only cancerous or not, also improve architecture and system improve large dataset.

In [9] Manasee Kurkure1, Anuradha Thakare2. They have used Genetic candidate Group search (GCGS) Algorithm and Naïve base classifier. This paper has some limitations are improve classification methods, low accuracy and large amount of dataset should be require for better result.

In [12] Suren Makajua, P.W.C. Prasad*a, Abeer Alsadoona, A. K. Singhb, A. El- chouemic. They have used water shaded Transform and SVM Classifier. This paper has some limitations are does not classify degree of nodule, not classify different stage and also improve the accuracy must be in- crease by pre-processing.

In [3] Mehdi Fatan Serj, Bahram LaviGa- briela Hoff and Domenec Puig Valls. They have described about DCNN for lung cancer detection approach and they have used DCNN classifier. This paper has some limitations are segmentation is not done, big feature vector and GPU is requiring to run the system.

In [15] F. Ghazvinian Zanjani1, A. Panteli1, S. Zinger1, F. van der Sommen1, T. Tan1B. Balluff2, D. R. N. Vos2, S. R. Ellis2, R. M. A. Heeren2, M. Lucas3, H. A. Marquering3I. Jansen3, C. D. Savci-Heijink3, D. M. de Bruin3 and P. H. N. de With1. They have describe about cancer detection in Mass Spectrometry image data using RNN (Recurrent Neural Network). Some limitation of this paper is improving architecture of RNN, Im- prove classification result and increase performance.

In [4] Moffy Vas, Amita Dessai. They have described about for detecting lung cancer methods which they have used are Morpho- logical, Harrick, GLCM and ANN. This paper has some limitations are Morphological segmentation require to change structuring element and they classify only two classes.

III. METHODOLOGY

A. Datasets [1]:

The data are a tiny subset of images from the cancer imaging archive. They consist of the middle slice of all CT images taken where valid age, modality, and contrast tags could be found. This results in 475 series from 69 different patients.

Data Header

- Images (DICOM, 18.3GB)
- Tissue Slide Images (web)
- Clinical Data (TXT)
- Genomics (web)

TCIA Archive Link – https://wiki.cancer imagingarchive.net/display/Public/TCGA- LUAD

B. Segmentation:

i. OTUS Thresholding [6]: Otsu method is one of the most successful methods for image thresholding. Converting a greyscale image to monochrome is a common image pro- cessing task. Otsu's method, named after its inventor Nobuyuki Otsu, is one of many binarization algorithms. In the simplest form, the algorithm returns a single intensity threshold that separate pixels into two classes, foreground and background.

ii. Watershed segmentation [1,8]: Watershed is a transformation de- fined on a grayscale image. The watershed transformation treats the image it operates upon like a topographic map, with the brightness of each point representing its height, and finds the lines that run along the tops of ridges. Change image into another image whose catchment basins are the objects you want to identify.

iii. Super Pixel Segmentation [12]: Separation of objects and regions of interest from the other parts of the image is used so that the image can be properly analysed. The image pixels are classified into anatomical region, such as muscles, bones and blood vessels or into pathological regions such as tissue deformities, multiple sclerosis and cancer based on its usefulness in a particular application

iv. Morphological based segmentation [4]: Converting the images to binary reduces computational complexity and storage issues and also is a pre-requisite for morphological segmentation of lungs.

- Morphological open operation,

$$A \circ B = (A \theta B) \oplus B$$

- Morphological closing operation,

$$A \bullet B = (A \oplus B) \theta B$$

v. PDE base segmentation [6]: PDE (Partial Differential Equation is also used for the purpose of segmentation. In this, active contour model and level set method are used and for filtering process, Anisotropic diffusion has been used.

C. Feature Extraction

TABLE I

SHAPE FEATURE

Area [6,7]	When the boundary points change along the Shape boundary, the area of the triangle formed by two successive boundary points and center of gravity also changes.
Perimeter [6,7]	The perimeter of a two-dimensional shape is the distance around the shape. It is found by adding up all the sides.
Eccentricity [6,7]	the eccentricity of a conic section is a non-negative real number that uniquely characterizes its shape. $e = \frac{\sin \beta}{\sin \alpha}, 0 < \alpha < 90^0, 0 \le \beta \le 90^0$ where β is the angle between the plane and the horizontal and α is the angle between the cone's slant generator and the horizontal.

Diameter [7]	A diameter of a circle is any straight-line segment that passes through the centre of the circle and whose endpoints lie on the circle.
Centroid [7]	the centroid or geometric centre of a plane figure is the arithmetic mean position of all the points in the figure. Informally, it is the point at which a cut-out of the shape could be perfectly balanced on the tip of a pin.
Mean Intensity [6,7]	The total number of individuals of a particular parasite species in a sample of a host species -^ number of infected individuals of the host species in the sample.

TABLE III

TEXTURE FEATURE

Haralick feature extraction [4]	Where Ng is the number of gray level, p is the normalized symmetric GLCM and p(i,j) Element of normalized GLCM, $Energy = \sum_i \sum_j p(i,j)^2$ Energy calculates the local uniformity of the gray levels in an image. Higher the similarity in pixels, higher is the energy value. Correlation= $\sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y}$ Where $\mu_x, \mu_y, \sigma_x, \sigma_y$ are the means and standard deviations of the GLCM. Correlation is a measure of linear dependency of gray intensity values in the co-occurrence matrix. $Variance = \sum_i \sum_j (i-\mu)^2 p(i,j)$ Variance feature measures the spread of intensity values of GLCM pixels about the mean. It is similar to entropy. $IDM = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$ Inverse difference moment (IDM) gives an account of the local homogeneity in the image. When the local gray level in an image is uniform, IDM is high. $DiffrenceEntropy = -\sum_{i=0}^{N_g-1} P_{(x-y)}(i) \log(P_{(x-y)}(i))$ $IMC1 = \sum \frac{HXY - HXY1}{\max\{HX, HY\}}$ IMC1 is the information coefficient of correlation I, where, $HXY = -\sum_i \sum_j p(i,j) \log(p(i,j))$ $HXY1 = -\sum_i \sum_j p(i,j) \log\{p_x(i) p_y(j)\}$ $HXY1 = -I \sum_i \sum_j p_x(i) p_y(j) \log\{p_x(i) p_y(j)\}$ $Contrast = \sum_i \sum_j (i-j)^2 P(i,j)$ Contrast indicates the intensity variations between the pixel under consideration and its neighbouring pixel. Larger contrast means larger variation.
Gabor filter [8]	In image processing Gabor filter is a linear filter used for Texture analysis, which means that is basically analyze whether there is any specific frequency content in image in specific direction in localized region around the point or region of analysis.
GLCM [1,13]	A gray level co-occurrence matrix (GLCM) contains information about the positions of pixels having similar gray level values. A co-occurrence matrix is a two-dimensional array, P, in which both the rows and the columns represent a set of possible image values. A gray level co-occurrence matrix (GLCM) contains information about the positions of pixels having similar gray level values.

D. Classification

1) Machine Learning [11]: Machine learning focuses on the development of computer programs that can Access data and use it learn for themselves. Need to try different features and classifier to achieve the best result. Quick training model. Machine learning is enabling analysis of massive quantities of data. For example, medical diagnosis, image processing, prediction, classification, learning association, Regression etc.
    - SVM [1,7]: Support vector machine is supervised machine learning algorithm which can be used for both Classification or regression challenges. It can solve linear and non-linear problem and work well for any practical problem. The idea of SVM is simple: The algorithm creates a line or a hyper-plane which separated the data into classes.
    - K-NN [8]: K- nearest neighbours is one of the simplest algorithms used in machine learning for regression and classification problem. KNN algorithm use data and classify new data point based on similarity measures. KNN is a supervised learning algorithm used for classification.

2) Deep learning [10]: Deep learning is a subset of machine learning where artificial neural networks Algorithms inspired by the human brain, learn from large amounts of data. Deep learning allows machines to solve complex problems even when using a Data set that is very diverse, unstructured and inter-connected. Computationally intensive model Deep learning applications are used in industries from automated driving to Medical devices.
    - CNN [2,3]: In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analysing visual imagery. CNNs are regularized version of multilayer perceptron. CNN are special type of Feed-Forward Artificial Neural Network that are generally used for image detection tasks. [10]
    - DNN [3]: A deep neural network is a neural network with a certain level of complexity, a neural network with more than two layers. Deep neural networks use sophisticated mathematical modelling to process data in complex ways. A deep neural network (DNN) is an artificial neural network (ANN) with multiple Layers between the input and output layers. The DNN finds the correct mathematical manipulation to turn the input into the output, whether it be a linear relationship or a non-linear relationship.
    - RNN [15]: RNN short for “Region Neural Network” A recurrent neural network (RNN) is a class of artificial neural networks where Connections between nodes form a directed graph along a temporal sequence. A recurrent neural network (RNN) is a type of artificial neural network commonly used in speech recognition and natural language processing (NLP). RNNs are used in deep learning and in the development of models that simulate the activity of neurons in the human brain.

IV. COMPARATIVE STUDY

TABLE IIIII

COMPARATIVE STUDY

Method	Advantage	Limitation
Segmentation		
OTSU [6]	-Easy to Implement. -Fast Processing.	-Work with image intensity only. -Gives false output with Low Contrast image.
Watershed [1,8]	-Provides closed	-Work with gray scale image

	contours. -Requires low computation time	only. - Over-segmentation refers to over-cutting that occurs.
Super Pixel [1,2]	-Covers smoothness constraints. -Works with edges more.	-Very high computational complexity. -Relatively poor boundary adherence performance
Morphological [4]	-Fast Processing. -Low Complex.	- It produces excessive over-segmentation. -Require Structuring Element.
PDE [6]	-It is implicit, is parameter-free, and provides a direct way to estimate the geometric properties of the evolving structure.	-High computational complexity. -Require memory.
Features		
Shape [7]	-Easy to implement, -Less Complex, -Less Time Consuming	-Works with Binary Image only.
Haarlick [8]	Computational accuracy of feature vectors is high, Classification accuracy is high	-Due to 13 features the computation of feature vectors is complex. -High Dimensionality
Gabor [8]	Achieves highest retrieval results. Support Orientation & Scaling	-Only consider Gray scale images. -Time Consuming.
GLCM [1,13]	Computation Time is Low. Supports all type of texture. Low memory Consumption.	-Works with Gray scale images
Machine Learning		
SVM [1,7]	-SVM is less complex. -Produce very accurate classifiers. -Less over fitting, -Robust to noise.	-SVM is binary classifier, to do a multi-class classification, pair-wise classifications can be used. -Computationally expensive, thus runs slow
KNN [8]	-Robust to noisy training data -Effective if the training data is large	-Distance based learning is not clear which type of distance to use and which attribute to use to produce the best result. -Computation cost is quite high
Deep Learning		
CNN [2,3]	High degree of	-Hard to tune

	non-linearity possible. -It is suitable for spatial data such as images.	parameters. -Takes time to build model.
DNN [3]	- DNN includes high feature compatibility. -Better accuracy.	-Some time GPU unit require.
RNN [15]	-RNN is suitable for temporal data, also called sequential data. - RNN includes less feature compatibility.	- Use Internal memory to process arbitrary sequences of inputs.

V. CONCLUSION

NSCLC Stage classification is requiring for early diagnosis. Past method is based on classify only lung cancer type. It would not provide stage information. In this research, we summarize different types of lung cancer stages for classification. For that different feature and deep learning approaches introduce new strategy combine shape and texture feature. For shape feature we will use segmentation based on clustering and after that deep classification for future stage prediction.

VI.REFERENCES

[1] J. Alam, S. Alam, and A. Hossan, "Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classified," Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018, pp. 1–4, 2018.

[2] E. Cengil and A. Çinar, "A Deep Learning Based Approach to Lung Cancer Identification," 2018 Int. Conf. Artif. Intell. Data Process. IDAP 2018, 2019.

[3] B. Lavi, "A Deep Convolutional Neural Network for Lung Cancer Diagnostic," pp. 1–10.

[4] M. Vas and A. Dessai, "Lung cancer detection system using lung CT image processing," 2017 Int. Conf. Comput. Commun. Control Autom., pp. 1–5, 2017.

[5] S. Moreno, M. Bonfante, E. Zurek, and H. S. Juan, "Study of medical image processing techniques applied to lung cancer," Iber. Conf. Inf. Syst. Technol. Cist., vol. 2019-June, no. June, pp. 1–6, 2019.

[6] P. Tripathi, S. Tyagi, and M. Nath, "A Comparative Analysis of Segmentation Techniques for Lung Cancer Detection," Pattern Recognit. Image Anal., vol. 29, no. 1, pp. 167–173, 2019.

[7] 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.

[8] P. Bhuvaneswari and A. B. Therese, "Detection of Cancer in Lung with K-NN Classification Using Genetic Algorithm," Procedia Mater. Sci., vol. 10, no. Cnt 2014, pp. 433–440, 2015.

[9] M. Kurkure and A. Thakare, "Classification of Stages of Lung Cancer using Genetic Candidate Group Search Approach," IOSR J. Comput. Eng., vol. 18, no. 05, pp. 07–13, 2016.

[10] R. Tekade and K. Rajeswari, "Lung Cancer Detection and Classification Using Deep Learning," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, no. 2, pp. 259–262, 2018.

[11] A. Shrestha and A. Mahmood, “Review of deep learning algorithms and architectures,” IEEE Access, vol. 7, pp. 53040–53065, 2019.

[12] A. Asuntha, A. Brindha, S. Indirani, and A. Srinivasan, “Lung cancer detection using SVM algorithm and optimization techniques,” J. Chem. Pharm. Sci., vol. 9, no. 4, pp. 3198–3203, 2016.

[13] M. Computing, N. Panpaliya, N. Tadas, S. Bobade, R. Aglawe, and A. Gudadhe, “a Survey on Early Detection and Prediction of Lung Cancer,” Int. J. Comput. Sci. Mob. Comput., vol. 4, no. 1, pp. 175–184, 2015.

[14] M. Kurkure and A. Thakare, “Lung cancer detection using genetic approach,” Proc. - 2nd Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2016, 2017.

[15] F. G. Zanjani et al., “Cancer Detection in Mass Spectrometry Imaging Data by Recurrent Neural Networks Eindhoven University of Technology, SPS-VCA, 5612 AJ Eindhoven, The Netherlands Maastricht Multimodal Molecular Imaging Institute, University of Maastricht, The Netherlan,” 2019 IEEE 16th Int. Symp. Biomed. Imaging (ISBI 2019), no. Isbi, pp. 674–678, 2019.

Cite this article as :

Jay Jawarkar, Nishit Solanki, Meet Vaishnav, Harsh Vichare, Dr. Sheshang Degadwala, " A Survey on Multistage lung cancer Detection and Classification", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 313-320, September-October 2020.
Available at
doi : https://doi.org/10.32628/CSEIT20631110
Journal URL : http://ijsrcseit.com/CSEIT20631110