Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020]
IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3

Lung Cancer Disease Diagnosis Using Machine Learning Approach

Swati Mukherjee
Student: Computer Science dept
G H Raisoni University,
Amravati, India.
Email : swati.raisoni@gmail.com

Prof . S. U. Bohra
Professor: Computer Science dept.
G H Raisoni University,
Amravati, India
Email : sneha.bohra@raisoni.net

2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.9315909

Abstract— **The analysis and study of lung diseases has been the most intriguing investigation zone of medical experts from early days to the present day. To address this concern, a diagnosis system like this can only help diminish the odds of getting risk to human live by early discovery of malignant growth. By and by a couple of structures are proposed and still an enormous number of them are still a hypothetical plan. In the ensuing philosophy, the performance of a neural network model is examined to address this issue of recognizing cancerous cells in image data, an average issue in therapeutic imaging applications. In an attempt to accomplish this task, a lung cancer identification framework is developed based on AI and deep neural system, wherein the methodology depends on supervised learning for which a better precision has been obtained, especially by using the deep learning mechanism. CNN classification is a game plan of lung tumor classification. The framework includes various methods, for instance, picture acquisition, pre-preparing, enhancement, segmentation, feature extraction, and neural framework identification. To put it concisely, machine learning approach can give an unprecedented opportunity to improve decision support in lung cancer treatment at low cost.**

Keywords—convolutional neural network(CNN); lung cancer disease;supervised learning ;

I. INTRODUCTION

Due of enormous pervasiveness of smoking and air contamination around the globe, lung malignancy is a lethal ailment and thus global problem in ongoing decades. Lung Cancer constitute 41% of cancer burden in India. It happens when cells in the lung mutate. They grow disorderly and cluster together to form a tumor.

Thus it is crucial to diagnose metastatic tumor at initial stage itself to ensure timely treatment and increase survival rate. However, the diagnosis process is time consuming and costly as it requires human intellect to make critical decisions. After exploring the elegance of machine learning in all other domains for image object detection, it is strongly believed that deep learning methods can contribute largely towards this pressing issue.

In this current examination, it has been proposed to use the profound neural network system to improve the chances of tumor identification. Here, supervised learning approaches are utilized with NSCLC Radio genomics lung malignancy CT (computed tomography) picture dataset. A framework is built that comprises of numerous steps, for example, image extraction, pre-processing, binarization, thresholding, segmentation, feature extraction. It is an application which will take input as the CT scan images and will predict the possibilities of the malignancy and its stages using deep learning. The proposed research work has dealt with DICOM dataset and utilized profound Convolutional neural system for accomplishing high precision.

II. OBJECTIVE AND SCOPE

1. Main goal is to build up a framework that helps the clinical specialists to cross confirm their analyzed results of predicted lung cancer.
2. As the existing diagnosis process is time consuming, laborious and costly, by and large this deep learning based tool can identify the tumor growth and predict stages.
3. Since this is automated tool based on image processing and AI, it minimizes human effort in predicting the presence of cancer cells from image.

III. PROBLEM STATEMENT

- Detection of lung cancer lesions using image data and not symptoms.
- It is a tool which will take input as the CT scan images and it will predict the possibilities of the disease and its stages using deep learning.
- To give higher exactness over past research.
- To give exceptional and most encouraging instrument that can help the specialists to identify the cancer at early stage.

IV. REVIEW OF LITERATURE

In this paper, YutongXie, [1] states that a multi-view information based collective (MV-KBC) deep model was used to isolate malignant tumor from normal lung nodules utilizing chest CT information. They used 9 KBC[1] sub-models to train the model. The model was tested on LIDC-IDRI data set and compared with the five modern classification approaches.

Qing Wu and Wenbing Zhao [2] indicated that An EDM AI calculation with vectored histogram can be used to distinguish SCLC for early malevolent malignancy identification.

LilikAnifah et.al [3] proposed the detection of lung cancer utilizing Artificial Neural Network Back-propogation based Gray Level Co-event Matrices (GLCM) features. The lung information is utilized from the Cancer imaging archive Database, comprised of 50 CT-pictures. The steps of this process are: image pre-processing, segmentation, feature extraction, and recognition of tumor growth using Neural Network Back-propagation method which has 3 layers. The result showed that framework[3] can differentiate between ordinary lung and lung malignancy with accuracy of over 80%.

Prof. AnuradhaDeshpande and DhaneshLokhande[4] focused on lung cancer prediction using image processing strategies followed by watershed segmentation and SVM.. In this combination procedure, the critical characteristics of various pictures are consolidated together to acquire the required data in a Fused Image format. CT picture examines the denser tissues and MRI filters the delicate tissues, so by joining pertinent data of the two pictures, proper data of melded picture is obtained. This procedure additionally enhances the quality of the melded picture.

Abbas Khosravi and Amin Khatami [6] examined that classification was initially difficult using High dimensional dataset but in this approach, classification was possible using autoencoders and deep learning [6].

V. PROPOSED SYSTEM APPROACH

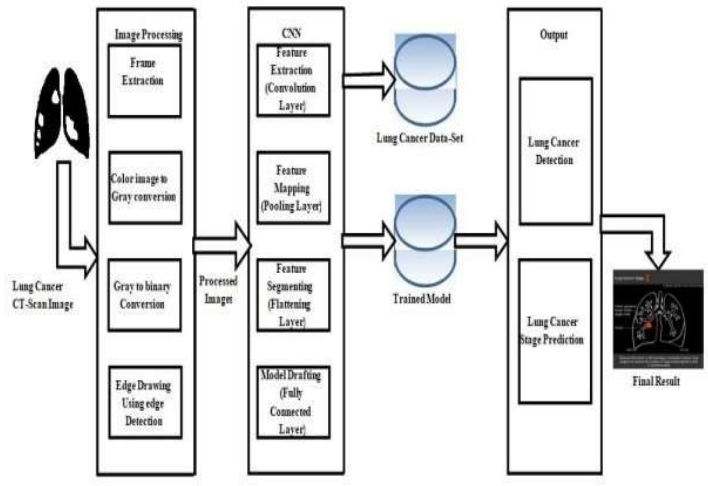


Fig. 1. System Architecture

This model comprise of following phases.

• Image Processing:-
A picture is comprised of RGB hues. In Image processing, some operations are performed on image in order to get an enhanced image or to extract some valuable data from it. Thus the input is usually the image and output may be image or certain properties associated with the image.

• Image Filtering :-
Filtering is a procedure to change or improve the picture, for example to show certain features or eliminate other features. It incorporates smoothing, honing, removing noise and edge upgrade. Picture can be filtered in frequency or spatial domain. It is a process in which value of any given pixel in output image is found out by applying some algorithm to values of pixels in neighborhood of corresponding input pixel.

• Feature Extraction:-
It includes extraction, by which certain features of interest within an image are detected and represented for further processing. It is a critical step as it marks the transition from pictorial to alphanumerical data representation.
In our framework, CNN is incorporated.

• Segmentation :-
It is a method of partitioning a digital image into multiple segments(set of pixels) and so is used to locate objects and boundaries viz. lines, curves etc in an image. All pixels in a region or segment share a common property. One simplest method is thresholding.

• Edge Detection :-
Edge Detection is most fundamental tool in image processing and computer vision that is used to identify points in digital image at which the image brightness changes sharply and has discontinuities. It is most well-known method to find significant discontinuities in intensity values precisely. Edges appear on boundary between two regions. Most of the shape information of an image is hidden in edges of the image and needs to be extracted. Algorithm utilized in our exploration is canny edge detection operator.

• Feature Recognition:-
A feature is region of a part with some interesting geometric properties. The purpose of feature recognition is to mathematically extract higher level features(manufacturing data) from lower level entities(e.g. surfaces, edges, curves). This is a pioneering technology that extracts features and their parameters from solid models. It is the ability to identify and group topological entities [7] such as faces in solid model , into functionally important features such as ribs, holes, slots.CNN identifies the relation and connection within information and learns (or is prepared) through experience, not from programming.

VI. MATHEMATICAL MODEL

- **System Description:** S= {I, F, O} S be the lung cancer detection and stage prediction system.

INPUT:

- F=F1,F2,F3...FN Functions to be execute modules
- **I**=C1,C2,C3... input of systems lung cancer images
- **O**=R1,R2,Rn prediction results in the form of stages
- I=result access by User

F:

- **F1**=Image processing applied on Lung Cancer images
- **F2**=feature extraction from images
- **F3**=Stage detection results

O:

- R1= model creation from training.
- R2= model based image testing

Success:

1. High accuracy achieved by using CT image dataset.
2. Any user can use this tool to get the result faster.

Failures:

1. Huge database may lead to more time consumption to get the required image information.
2. Hardware failure.
3. Software failure.

Mathematical Model in Equation format Notation

Where,

- M= Set of all entities.
- LCIT1= Lung cancer images type 1
- LCIT2= Lung cancer images type 2
- LCITN= Lung cancer images type N
- TLCI=Total Lung cancer images

We calculate total number of images by following Equation:

Total number lung cancer images = Total number lung cancer images type 1+ = Total number lung cancer images type 2+…..+ = Total number lung cancer images type N

$$\sum \textbf{TLCI} = \sum \textbf{LCIT1} + \sum \textbf{LCIT2} + \ldots + \sum \textbf{LCITN} \ldots \text{Equation(1)}$$

VII. ALGORITHM

Convolutional Neural Network (CNN)

Detailed steps of the convolution network are mentioned below.

• Image handling utilizing Convolutional neural systems (CNN) has been utilized in different fields, for example, facial recognition, analyzing documents, historic and environmental collections, understanding climate, drug discovery, video analysis, advertising etc.

• CT scan images are considered for this ongoing research work. Due to high complexity of lung cancer pathology images, predicting patient outcome from lung cancer is still very challenging and will require large amount of data for model developments. So, current focus is on developing patient outcome prediction models based on image features extracted based on deep learning classification using CT scan images.

• Eventually, using such deep learning techniques, it makes easy to process various formats of data, such as imaging and genomic features.

• Processing of pictures with CNN includes various procedures, for example,

1. Image pre processing includes steps such as reading image, resize, remove noise and other morphological operations. The objective of pictures pre-dealing with CNN is improving, restoring or redoing pictures , removing unwanted distortions and enhancement.
2. Feature extraction incorporates eliminating different entities and selecting more relevant ones and thus combining to a new reduced set of features.
   It helps in compacting the image followed by isolating geometric traits (edges, corners, and joints), facial features, etc.
3. Segmentation is a division of an image into locales.
4. Recognition includes the identifying and detecting an object or feature in image.

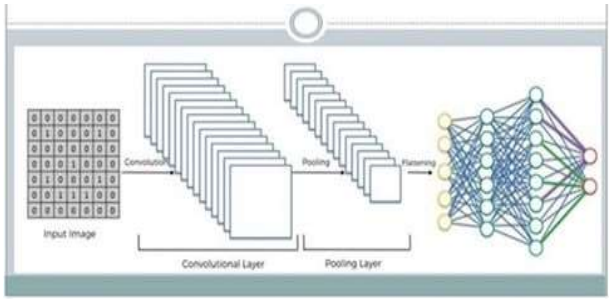


Fig. 2. Block Diagram CNN

CNN consist of following 4 layers as shown in Fig.2

- Convolutional Layer
- Pooling
- Flattening
- Fully Connected layer

1. Convolutional Layer

In convolution layer, we extract different features pixel wise by using feature detectors/kernels. Perform numerous convolutions on input, where each operation uses a different filter. This results in different feature maps. Later, all feature maps are taken and put together to give final output of the convolution layer.

2. Pooling

This decreases the training time and solves the problem of over fitting. Max Pooling extracts out the highest pixel value out of a feature that has to be extracted.

3. Flattening

In flattening, we arrange the pooled feature into a single vector/column as an input for next layer (convert our 3D data to 1D).

4. Fully Connected Layer

In fully connected layers, every neuron in one layer is connected to every neuron in another layer. It is similar to multi layer perceptron(MLP). This layer takes the inputs from the previous feature analysis step and adds weights to predict the correct label. Result of this is the final classification decision.

Convolutional Neural Network Approach (CNN)

Step 1- Input lung cancer image

Step 2- Image processed by using open-cv

Step 3- Feature Extraction from images

Step 4- Machine model generation

Step 5- Lung cancer stage classification

Step 6- Stage detection

Tensor Flow

It is an open source math library especially for numerical computation using dataflow graphs. TensorFlow[7] can train and run deep neural networks for classification, image recognition and prediction. In this current effort, CNN is implemented using TensorFlow framework.

Python

The python has rich set of predefined machine learning library. There are many in-built python functions that are used in our algorithms. It is versatile, object oriented and high level programming language.

Open-CV

Open-cv is image processing library used for image related functionalities. The lung cancer image can be read using this library and pixel weights is extracted from images. OpenCV-Python utilizes Numpy library, which is most advantageous for numerical operations. So in our project, images are transformed, read and displayed using opencv library.

Image Processing Image processing phase has main steps like Gaussian filtering, binarization, smoothing of images, and edge detection etc .

VIII. COMPARATIVE RESULTS

Table 1. Lung cancer stage evaluation

Primary Tumor(T)	Criteria
T1	< 3cm in diam; T 1a <= 2cm; T 1b > 2cm <= 3cm
T2	> 3cm <= 5cm; T 2a > 3cm <= 4cm; T 2b > 4cm <= 5cm
T3	>5cm<=7cm
T4	Any Size greater than above

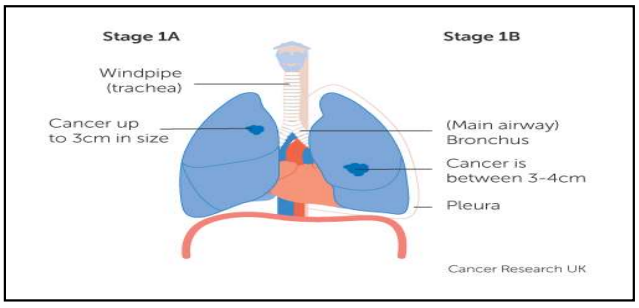


Fig. 3. Lung Cancer nodules in Stage 1

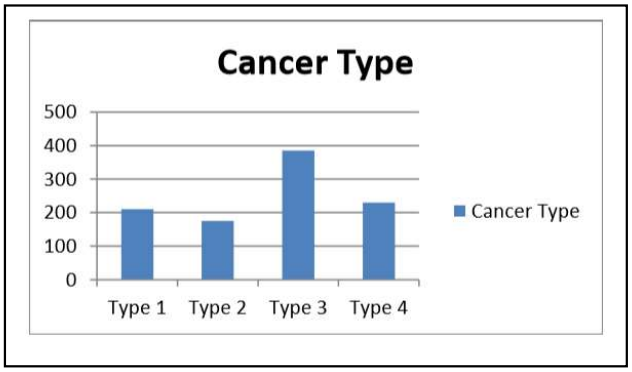
In our experimental setup, in Table no.2, it shows total number of 1000 images of lung cancer. Our project contains mainly Lung cancer images type 1, Lung cancer images type 2, Lung cancer images type 3, and Lung cancer images type 4.

This project consists of 210 number of images of lung cancer type 1, 175 number of images of lung cancer type 2, 385 number of images of lung cancer type 3 and 230 number of images of lung cancer type 4.

Table 2. Model Testing of Image data

Sr No.	Types	Number of images
1	Lung cancer testing dataset images type 1	210
2	Lung cancer testing dataset images type 2	175
3	Lung cancer testing dataset images type 3	385
4	Lung cancer images testing dataset type 4	230

From above data, we created the Graph 1 as below, the total numbers of images of lung cancer type 1 were 210, total numbers of images of lung cancer type 2 were 175, total numbers of images of lung cancer type 3 were 385 and total number of images of lung cancer type 3 were 230.



Graph 1. Total number of cancer type

IX. CONCLUSION

We studied various relevant work done on this disease and attempted to use best image pre processing methods, feature extraction and deep learning mechanism to get more accurate prediction results.

Ultimately, profound neural system approach is utilized to achieve more accurate precision in discovery of lung malignancy and precisely anticipate stages. Likewise, we could demonstrate that utilization of AI can possibly fundamentally distinguish and group low populace.

X. FUTURE WORK

More number of images can be considered like X ray, CT, MRI, PET that will bring about more accuracy, thereby helping the medical practitioners to offer quick prophylaxis at low cost.

As this disease has detrimental economic impact, further analysis can be done to identify the knowledge gaps in disease control and diagnosis methods which can facilitate development of vaccine or other control measures.

Acknowledgment

This work is supported for lung cancer disease diagnosis using machine learning approach research field in India. Authors are thankful to Faculty of Computer Engineering Dept, G H Raisoni University, Amravati for providing the facility to carry out the research work

References

[1] YutongXie,, "Knowledge-based Collaborative Deep Learning for Benign Malignant Lung Nodule Classification on Chest CT" ,2018, IEEE .

[2] Vijayakumar, T."Classification of Brain Cancer Type using Machine Learning." *Journal of Artificial Intelligence* 1,no.02 (2019): 105-113.

[3] LilikAnifah, Haryanto, RinaHarimurti, "Cancer lung detection on CT Scan image using ANN backpropagation based gray level co occurrence matrix feature." 978-1-5386-3172-0/17/ 2017 IEEE .

[4] Prof. Anuradha Deshpande, Dhanesh Lokhande, "Lung cancer detection with fusion of CT and MRI image using image processing." (IJARCET) Volume 4 Issue 3, March 2015 .

[5] RachidSammouda, "Segmentation and analysis of CT chest images for early lung cancer detection." Global Summit on Computer & Information Technology 978-1-5090-2659-3/17 2017 IEEE.

[6] Abbas Khosravi, Amin Khatami, "Lung cancer classification using deep learned features on low population dataset." Canadian Conference on Electrical and Computer Engineering (CCECE) 978-1-5090-5538-8/17 2017 IEEE .

[7] https://en.wikipedia.org.

[8] https:// www.icmr.nic.in