Identification of Lung Cancer using Image Processing and Convolutional Neural Networks

Devanshu Pipalia[1], Sunny Shah[2], Shriya Pisal[3], Prof. Jaya Zalte[*]

[1,2,3]*B.E student, Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India*

[*]*Professor, Department of Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai, Maharashtra, India*

---***---

Abstract - *Biomedical Image Processing is the latest technology used in medical field for the early detection of cancer. Similarly, Artificial Intelligence can also be used in the medical field to diagnose cancers at an early stage. The inputs for image processing are going to be Computed Tomography (CT Scans) of lungs of the patients which will be taken from the open database, Lung Image Database Consortium (LIDC). During the pre-processing step of images, the RGB images are converted to gray-scale images as RGB images are too complex to process. Gray-scale image are further converted into Binary images. After the Processing stage, the CT scan images become much more efficient and refined for the further usage. These are inputs for the Convolution Neural Network. CNN also includes steps like Convolution Filtering, Max Pooling filtering which help in training the data to accurately predict whether the CT Scan image is cancerous (malignant) or non-cancerous (benign). Deep Learning is a latest branch of Artificial Intelligence which will help to improve the performance of CNN based systems. The proposed system will also consider the time taken for processing and the delay caused for the efficiency of cancer detection.*

***Key Words*: CNN, Deep Learning, LIDC, image processing, CT scan, watershed segmentation**

1. INTRODUCTION

Lung cancer is the most common cancer among men today, and is the third most common cancer among women. Tobacco and smoking causes 85 per cent of cases of lung cancer. Approximately 20 per cent of all cancer deaths are due to lung cancer. 2.1 million new cases were reported in 2018, and 1.8 million deaths from lung cancer. The five-year survival rate for tumors identified at an early stage is 56 percent, i.e. if the cancer is detected at an early stage, then therapies will avoid tumor mortality and spread for at least 5 years, if not diagnosed more or more thoroughly. The mortality rate among individuals can be decreased by 20 per cent if the tumor is identified at an early stage. As the infection spreads, cancer cells begin to grow, and tumor size gradually increases. Much of the cancer patient community has an economically poor history and cannot afford healthcare or training, let alone consultation and diagnosis. Artificial Intelligence can play an important role in a population where a computerized system will replace expensive healthcare.

Image Processing is the process of transforming an image to digital form and performing some operations. This is done with the aim of getting an improved picture or collecting any useful information. It involves steps such as collecting images from sources, eliminating noise, enhancing and segmenting. The RGB images are converted to grayscale and then to binary during Image processing. This will help to improve image, and the filters will eliminate noise. It will remove the Blurred effect if any. That will improve the image quality [13]. The Image Processing Toolbox is used in MATLAB to perform the Image Processing. For each stage of image processing several different algorithms are possible [14].

Deep Learning is also used to identify CT Scan Images as cancerous / non-cancerous. Within Convolution Neural Networks the process of extracting features is such that features are described and determined by the algorithm itself. Input and an output label shall be provided during the training stage. The algorithm analyzes the characteristics / patterns on the basis of the data given and for training data, constructs a set of parameters and extracts features[15]. The new data can be evaluated on the basis of the computations for predicting a correct performance. Convolution Neural Networks consists of an input and output layer, and several layers that are hidden. The input layers accept inputs, and the number of output layers determines the corresponding number of outputs. Convolution layers are used to define the parameters and features. The pooling layers brought the computations together with equal permutation. By assigning a common value to a set of matrix pixels, the convolution filter can generate a spatially dense output. The production for that picture is determined by these values.

2. Related Work

In 2017, Lei Fan and his group of researchers from China took advantage of a deep learning algorithm to detect CAD lung cancer. In this paper, instead of applying image processing to the CT scanning images, they are fed directly as input to the convolution layers consisting of two convolution layers, two max-pooling layers, one completely connected layer and one output layer [1]. The machine provided 67.7 per cent accuracy. It implies that in the same number of input samples, Support vector machines have lower classification accuracy than 3D convolutional neural network.

The authors Pooja R. Katre and Anuradha Thakare discussed the different image processing methods used to detect lung cancer in 2017. Our key emphasis was to use a systematic approach to eliminating and improving the noise, for which a Median filter process was used. The reason behind using the median filter is that without blurring the image, it eliminates noise and also preserves the edges of regions [2]. It helps remove the noise of salt and pepper from the image. CT scanning images are input to the device, after which image processing is applied. In addition, it undergoes extraction stage of feature where field, perimeter, image excentricity is determined. All of these features help identify the tumor size so that the treatment process can take place as soon as possible.

Md., May 2015. Badrul Alam Miah and Mohammad Abu Yousuf have developed an early-detection and diagnosis CAD system based on the Neural Network. The results are focused on ANN and Fuzzy Clustering, IP, Cuvelet, Transform, Bayesian Multinomial Algorithm, Backpropagation, and Gray-coefficient. The main objective is to build a system with rotation and scaling that is fast and robust, more accurate [3]. The author used a 300-image dataset obtained from hospitals. Image generation, processing, binarization, segmentation, selection of features, and classification of the neural networks are different steps in the system involved.

Suren Makaju, P.W.C Prasad, Abeer Alsadoon and A.K. Singh during December 2017 worked on the CAD lung cancer program with the primary focus being on CT Scan Images. For this work, they regarded CT scan images as the best input data [4]. The initial model used noise reduction algorithms before Image Processing took place. For segmentation the model uses a watershed algorithm that is the same as the current system. Before classification it promotes a well-defined extraction of the function using SVM. The author used images from the LIDC dataset where the system provides 92 per cent accuracy.

In 2012, Mokhled S. Al Tarawneh published a comparison paper between different Image Processing methods and the algorithms they used for the lung cancer detection CAD method. The paper aims at detecting features for comparison of images with different processing techniques [1]. It is a three-step process starting with image enhancement that is used to improve the image quality, segmentation includes methods such as Thresholding and Watershed where the latter provides better segmentation quality and the binarization approach is used to extract features.

In 2017 Deep Prakash Kaucha and others suggested a method using the CAD system to increase precision, sensitivity and specificity in LCD. This maximizes the output at various points from the different techniques [9]. This uses K means for clustering, Gray Level Co-occurrence Matrix (GLSM) for extraction of elements, and classification backpropagation. The program suggested uses the LIDC database and is using MATLAB.

Prof. L.M. Deshpande and Mr. Vijay A. Gajdhane obtained photographs of CT scans from various hospitals for the identification of lung cancer using Image Processing. With the aid of MATLAB, they studied noise reducing filters, thresholding, segmentation and extraction of features. The images were passed through smoothing, enhancement, and then segmented with watershed segmentation as part of image processing, before the images were fed into convolution neural networks [11].

3. Proposed Solution

Convolution Neural Networks, which has better performance compared to other algorithms like SVM, ANN and Naïve Bayes, is used by the proposed system. Image Processing is performed on the data in order to improve its accuracy and to remove the Region of Interest from the input.

LIDC is a very critical biomedical access database which is well maintained. This program focuses on improving CNN-based system efficiency to identify malignant and benign cancer tumors for early cancer detection.

The performance of CNN-based systems can be enhanced by enhancing the quality of inputs or by offering better training or both. The established CNN network also plays a significant role in improving outcomes. More layers mean better preparation, and a more detailed computation on the given image.

The input images are pre-processed to enhance data quality that is sent to the CNN system. Preprocessing involves binarization, normalization and the elimination of noise from the following steps. That image with zero center normalization is normalized to 256* 256 pixels. This will bring the dataset to uniformity.

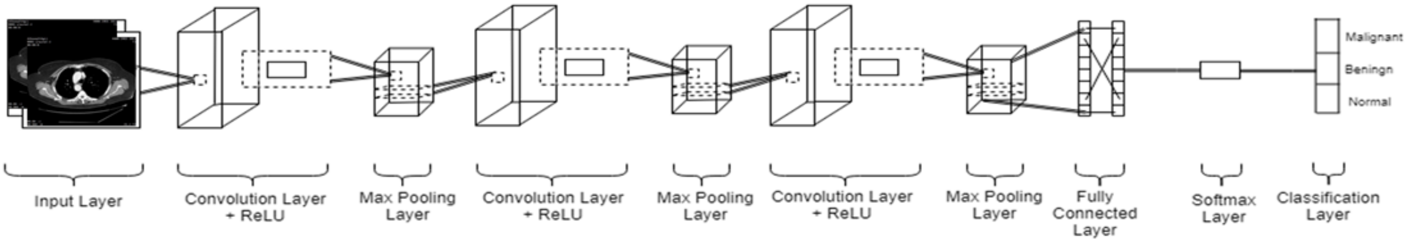

Fig 1: Architecture of CNN

3.1 Preprocessing

The input images are first transformed to grayscale, and then to binary image. Processing binary images is simpler than processing the RGB images. The CT scan images are mostly in grayscale by nature, but the conversion helps remove any unrecognized RGB portion. For this reason the function rgb2gray is used in Matlab. The binarization process means either a black or a white value (0 or 1) is assigned to each pixel of the image. Throughout Matlab, binarisation uses the function im2bw. To provide a uniformed set of data, Normalization is used. The imresize method is used to provide a uniform size of pixels. The noise present in the photos (most common is salt and pepper noise) may be mistaken by the network for a tumour. Median filters are used for the reduction of noise and picture smoothening to prevent these common mistakes. Median filter is used because it's more effective and eliminates the noise without distorting the edges of the image. Median filters are used in Matlab with the feature medfilt2. Then, these processed images are fed to the CNN network as data.

3.2 Convolutional Neural Network

Previous implementation of 3D CNN Networks used a 3D CNN architecture to create two sets of attribute maps, including two separate convolution layers. The second layer contains two different layers of max-pooling applied to the feature maps. After the max-pooling layer a convolution layer gives a collection of resized maps of features. For each function map a second max-pooling layer is added. It helps connect Layer of Convolution to multiple data frames. The final layers are fully connected layers and activation of the drop-out layer ReLU in each layer of the architecture is necessary. This whole architecture is connected by a complete, closing layer and a softmax layer with a sustained learning rate. The architecture provides 67.7 per cent accuracy.

The CanNet architecture uses an input layer to construct a linear volume in which images are concatenated to 3D. Two Convolution layers, a max-pooling and a full-connected layer and an output layer are followed to the input layer. The first layer of convolution generates 78 features, the second layer of convolution identifies patterns in features to learn feature from the previous layer in a hierarchical way. Both layers are accompanied by a layer of the ReLU which fixes all negative activations to zero. Max-pooling layer reduces data size by reducing data dimensions. A dropout layer allows randomly to prevent neurons in the CNN network to allow the checking of new data on different neurons and to reduce the effect of prior testing on new test data in the event of similarities. The fully connected layer reduces the architecture to two neurons, one being benign and malignant for both the desired outputs. CanNet architecture offers 76 percent accuracy.

Information to the CNN architecture in the proposed method will be individual slices of CT scan images in order to prevent a 3D concatenation and will analyze individual examples of each information to determine the existence of cancer tumor input. The architecture consists of 13 layers, since additional adjacent layers reduce the chances of performance errors. The layer list reads as follows:

1) imageInputLayer([227 227 3])
2) convolution2dLayer(5,20)
3) reluLayer
4) maxPooling2dLayer(2,'stride',2)
5) convolution2dLayer(5,20)
6) reluLayer
7) maxPooling2dLayer(2,'stride',2)
8) convolution2dLayer(5,20)
9) reluLayer
10) maxPooling2dLayer(2,'stride',2)
11) fullyConnectedLayer(3)
12) softmaxLayer
13) classificationLayer()

Input Layer provides the image input for the CNN network. The data to the network is a pre-processed image with normalization of the zero-center and empty transformations. With 3 forms of input data the image size is reduced to 227 * 227.

Convolution Layer employs two parameters; filter size and filter number. A 2-D convolution layer is added that converts the filters through the image, both vertically and horizontally. It calculates the weights and input point product for each function and adds a default value bias expression.

Rectified Linear Unit (ReLU) Layer applies a total function f(x)=max(x,0) to the matrix after convolution of the transformed picture. It sets all of the negative values in the matrix dot products to 0. All other principles remain unaltered. It increases the speed of network training by removing negative activations in the gradient, thus preventing complex negative computations.

After activation of ReLU, max-pooling layer with pool size[2,2] and stride[3,3] are established. This layer divides the input matrix into sub-regions of rectangular pooling, and calculates the limit for each region. This conducts down-sampling and renders the features in each sub-region an abstract representation. It reduces the pixel count which reduces the input parameter calculation. A 2 * 2 move prevents sub-regions from overlapping. In the sequence – Convolution Layer – > ReLU Layer – > Max-Pooling Layer, the layers Convolution, ReLU and Max-Pooling are added thrice.

The series of Convolution, ReLU, and Max-Pooling layers is implemented in complete Linked Layer. A valid, low-dimensional invariant feature space is the input to a fully connected layer. A non-linear combination of features is used to achieve the fully connected layer. It holds an input feature vector, which is necessary for classification or regression and categorization. The input size to board, ' 3 ' determines the three desired output sizes. A ' tf ' transfer function demonstrates the relation between input and output in the layer. Fully Connected layer provides the network with end to end instruction.

Softmax Layer has a softmax feature added to the output. The softmax function applies distribution of probability to the vector of features generated in a fully connected layer. Using probability distribution it maps the feature vector over a predicted output class.

Classification Layer will allocate each of the mutually exclusive output classes with a probability value derived from the softmax layer. It uses a cross-entropy function to assign an output class to each input (the value of the probability distribution).

3.3 System Flow Diagram

The CT Scan images are stored as a dataset separated into images that are malignant, beningn and normal. 70% of the dataset will be pre-processed and stored as a training dataset. Pre-processing includes conversion of the grayscale, reduction of noise and segmentation. This dataset is equipped with its labels within the CNN Neural Network. The remaining 30 percent of images are used as a repository for research. Images from the research dataset are pre-processed and forwarded for classification to the neural network. The end result is normal, beningn or malignant case detection along with edge-detected tumor using watershed algorithms.

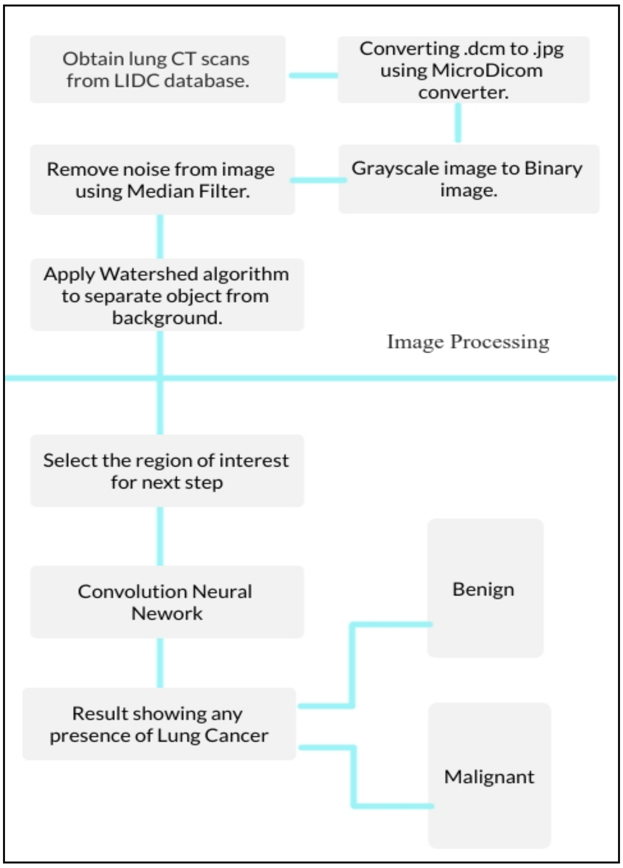

Fig 2: System Flow Diagram

4. Implementation

4.1 Implementation in MATLAB

The proposed system is implemented using MATLAB. It is a high-performance computing environment which is used in computer technology and various other fields for research, development, and study. The LIDC collection consists of images in DICOM format that is translated to.jpeg using the converter program for MicroDicom. The image pre-processing steps are applied to the images in MATLAB, and then transferred to another folder. The Deep-Learning Toolbox available on MATLAB is used for help relating to Deep Learning and CNN Neural Networks. The GUIDE; GUI Designing Framework can be used to design user interface in MATLAB; It is also possible to easily map two-dimensional or three-dimensional graphs on MATLAB too.

The proposed system uses malignant, benign and normal form of images as input. These images are initially in.dicom format, and are then converted to the format of .jpeg or .png. 250 images of each class are used for pre-processing and are stored in.png format as a pre-processed archive. Those photos are used by the CNN network for training purposes. After that the remaining images can be checked of each class in the dataset. Machine performance is either one of normal, benign, or malignant categories. The watershed algorithm used here displays tumor edges in a benign or malignant image. Confusion matrix can be used to measure the accuracy. In MATLAB version 2018b, graph plots are allowed to display the levels of training for the CNN network.

5. Results and Analysis

A total of 910 images were taken from LIDC for implementation in the framework proposed. Photos that had been labeled regular were 257, 331 beningn and 322 malignant. The CNN network has pre-processed and educated a total of 210 photographs of all 3 groups. The rest of the images have been untrained and checked to obtain the data. Of the 281 images in total, 47 images were regular, 121 images were beningn, and 113 images were classified as malignant class. Using uncertainty matrix the precision of the test is computed. Of the normal images, 10 false negatives were detected, and 9 images of other false positives were detected. From the beningn images, 10 false negatives were detected, and 8 false positives were detected. And from the malignant photos six false positives were detected and seven false negatives were detected.

6. Future Scope

The data set from the LIDC is a static dataset. Maintaining a complex real-time database can help monitor the changes occurring in various cases of lung cancer over a period of time. The increasing number of Convolutions will boost the results of this method in Deep Learning but it will also affect the delay in efficiency and performance. Sample size can be increased for training neural network for better GPU performance. The program may be checked using different databases such as the RIDER or TCIA database. In the future this system's overall accuracy can be improved for Deep Learning.

7. References

[1] 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..

[2] P. R. Katre and A. Thakare, “Detection of Lung ancer Stages using Image Processing and Data Classification Techniques”, 2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, 2017, pg. 402-404.

[3] 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.

[4] 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.

[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] 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.

[7] 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.

[8] R. Golan et. al, "Lung nodule detection in CT images using deep convolutional neural networks," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 243-250

[9] W. Alakwaa et. al, “Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN).” 2017 International Journal of Advanced Computer Science and Applications (IJACSA), volume 8, no. 8, pg. 109-117.

[10] S. Prakash et. al, “An Automated System for the Detection of Lung Cancer in CT data at Early Stages: Review,” 2017 International Journal of Control Theory and Applications, pg. 133-144.

[11]https://en.wikipedia.org/wiki/Median_filter , last accessed on 01/10/2018

[12] https://in.mathworks.com/help/images/marker-controlled-watershed- segmentation.html, last accessed on 19/08/2018

[13] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[14] https://www.wikipedia.com/lung_cancer, last accessed on 29/01/2019.

[15]https://en.wikipedia.org/wiki/Sensitivity_and_specificity, last accessed on 14/02/2019.

[16]https://in.mathworks.com/help/deeplearning/index.html,last accessed on 14/02/2019