Prediction Lung Cancer– In Machine Learning Perspective

1st Nikita Banerjee
Computer Science And Engineering
College of Engineering and Technology
Bhubaneswar,India
nikitabanerjee1994@gmail.com

2nd Subhalaxmi Das
Computer Science And Engineering
College of Engineering and Technology
Bhubaneswar, India
sdascse@cet.edu.in

Abstract— **Past years have experienced increasing mortality rate due to lung cancer and thus it becomes crucial to predict whether the tumor has transformed to cancer or not, if the prediction is made at an early stage then many lives can be saved and accurate prediction also can help the doctors start their treatment. Computed tomography plays a vital role in ensuring the condition of tumor that by checking the size of tumor, location of tumor, etc. In this paper, we have proposed a framework for prediction of cancer at an early stage so that many lives that are in an endangered situation could be revived. Basically, our focus is on two domains of computer science that is Digital Image Processing acronymed DIP and Machine Learning. Digital image processing is well-known for the phase of preprocessing the image. In the further stage, the pre-processed image is exposed to segmentation phase and then the segmented image is passed for feature extraction and finally the extracted features are trained using machine learning classification algorithms like SVM (Support Vector Machines), Random Forest, ANN (Artificial Neural Network) . Based on the classification results obtained, prediction is made whether the tumor is benign or malignant. The inevitable parameters such as accuracy, Recall and precision are calculated for determining which algorithm has the highest predictive accuracy.**

Keywords—Lung Cancer, Edge detection, Segmentation, SVM, Random Forest, ANN

I. INTRODUCTION

With the rapid increase in population rate, the rate of diseases like cancer, chikungunya, cholera etc., are also increasing. Among all of them, cancer is becoming a common cause of death. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow older or become damaged, they die, and new cells take their place. When cancer cells develop, however, this orderly process breaks down. As cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumor. This tumor starts spreading to different of body.

Tumors are of two types benign and malignant where benign (non-cancerous) is the mass of cell which lack in ability to spread to other part of the body and malignant (cancerous) is the growth of cell which has ability to spread in other part of body this spreading of infection is called metastasis. There is various type of cancer like Lung cancer, leukemia, and colon cancer etc. The incidence of lung cancer has significantly increased from the early 19th century. There is various cause of lung cancer like smoking, exposure to radon gas, secondhand smoking, and exposure to asbestos etc. Lung cancer is of two type small cell lung cancer (SCLC) and non small cell lung cancer (NSCLC). Non-small cell lung cancer is more common than SCLC and it generally grows and spreads more slowly. SCLC is almost related with smoking and grows more quickly and form large tumors that can spread widely through the body. They often start in the bronchi near the center of the chest. Lung cancer death rate is related to total amount of cigarette smoked. [1] Symptoms that may suggest lung cancer include:

- dyspnoea (shortness of breath with activity),
- haemoptysis (coughing up blood),
- chronic coughing or change in regular coughing pattern,
- wheezing, chest pain or pain in the abdomen, cachexia (weight loss, fatigue, and loss of appetite),
- dysphonia (hoarse voice),
- clubbing of the fingernails (uncommon),
- dysphasia (difficulty swallowing),
- Pain in shoulder, chest, arm, Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss [1]

To diagnose lung cancer various techniques are used like chest X-Ray, Computed Tomography (CT scan), MRI (magnetic resonance imaging) through which doctor can decide the location of tumor based on that treatments are given. Now it is important that the disease diagnose should be done in early stage so that many life's can be saved. As the medical images are full of noise and due to the present of noise it becomes very difficult for prediction. So, for that reason image processing technique will be applied on the medical image for pre-processing and then on the pre-processed image machine learning algorithm is implemented for predicting lung cancer.

II. ENABLING TERMINOLOGY

A. Segmentation

The objective of lung image segmentation is to extract the size of lung parenchyma from the preprocessed image and to remove windpipe, tubular branches, alveoli, and muscles from the image to give more accuracy and to reduce the complication while doing feature extraction. There is

numerous ways for segmentation like active contour model, Edge detection, Watershed segmentation etc. For Example for segmenting a lung CT scan image the steps are (1) apply edge detection on the preprocessed image (2) Then apply threshold to the edge so that if the image edge intensity is less than threshold will remove and the intensity which is more than the threshold will be considered. (3) Then morphological technique will be applied like morphological closing or opening. (4) After this step we will apply morphological segmentation to get the volume of the lung. [2]

B. Machine Learning Context to LungCancer

Machine learning is used for classify the tumor whether the tumor is benign or malignant. The algorithms in context to Lung Cancer Prediction are as follows:

1. Support Vector Machine- SVM help in prediction of cancer by separating the dataset into two classes by using kernel functions as the image data are high dimensional. As the images are arranged in non linear manner so it will plot the image in 3-D plane by using kernel function like polynomial kernel, Gaussian kernel, radial basis function etc, and separate the class using a hyper plane. For example if we take a CT scan of lung the image will be first pre-processed ad then the pre-processed image is trained by using RBF kernal while training the image are labeled as 1 and 2 for normal and abnormal tumor after training and testing it form a confusion matrix which show the prediction in two form classification and misclassification based on the classification table we can generate accuracy of our prediction.[3]
2. Artificial Neural Network- Artificial neural network is a concept generated from biological neural network. A multilayer feed forward neural network consists of input layer, hidden layer, Output layer. The image is first inserted into input layer is forwarded to calculate the activation value and then at output layer activation function is calculated and aggregated to get O(x), the difference between O(x) and desire output that is error is calculated using Backpropagation algorithm where some weight is assigned and the error deviates backward for optimal error value.[4]
3. Random Forest- Random is a collection of decision tree. It use the concept of bagging , it can handle many numbers of variable without deleting any variable.

III. SUMMARY OF LITERATURE SURVEY

Many works has already been proposed for prediction of cancer by various researchers among then Palani et al., [5] has proposed IoT based predictive modeling by using fuzzy C mean clustering for segmentation and incremental classification algorithm using association rule mining and decision tree for classification for classifying the tumor sets and based on the output generated by incremental classification model convolutional neural network has been applied with other features for predicting benign or malignant.

Lynch et al., [6] Various machine learning algorithm are implemented for predicting the survivability rate of person, performance is measured based on root mean square error. Each model is trained using 10-fold cross validation, as the parameters are preprocessed by assigning default value so cross validation is used for avoiding over fitting.

FENWA et al., [3] proposed a model whether feature like contrast, brightness from the image dataset is extracted using texture based feature extraction and on that two type of machine learning algorithm are applied one is artificial neural network another one is support vector machine and then performance has been evaluated on both the algorithm to compare which algorithm is giving more accuracy.

Öztürk et al., [7] proposed a model where a five type of feature extraction techniques were used in individual classification algorithm to predict at which features extraction technique which machine learning algorithm is giving more accuracy.

Jin et al., [8] proposed a model where the original image is first converted into binary image the erosion and dilution has been operated on that image after that image has been segmented on the segmented image region of interest extraction is applied to identify volume or size of the tumor and after extraction convolutional neural network is applied with softmax classification layer to recognize the tumor is cancerous or not.

Sumathipala et al., [9] proposed a model where the image data are taken from LIDC-IDRI, after collecting the image data image filtration has been implemented, filtration is done based on the patient who went through biopsy and module level is equal to 30 and then images whose module level is equal to 30 is segmented and then Logistic regression and random forest has been applied for prediction.

IV. PROPOSED FRAMEWORK FOR CANCER PREDICTION

Based on the literature survey a novel model has been proposed which consist of pre-processing block, segmentation block, feature extraction block and then classification block. In prediction of cancer CT scan report is basically used. But CT scan report is full of noise which cannot be seen by human eye for that reason various digital image processing plays a important role to get a noise free image. Digital image processing is the process where the analysis and manipulation of image is used to extract some useful information from the image. Digital image processing involve various step like image pre-processing where we can enhance the image using histogram equalization, spatial filter etc. Then image restoration can be done where various kind of noise like salt and pepper noise, Gaussian noise etc are applied and filter like median filter, mean filter can be applied on the pre-processed image. After that color conversions is applied only if the image is colored image then convert it to gray level. Fig. 1 shows the proposed novel framework.

Image segmentation is a process which divides the image into several segment based on the pixel, once the image segmentation is over the feature extraction can be applied. Feature extraction is a type of dimensionality reduction

where a set of raw data is reduced to more manageable group image data for extracting the feature like region and texture. After extracting the feature different machine learning technique is used to classify the image.

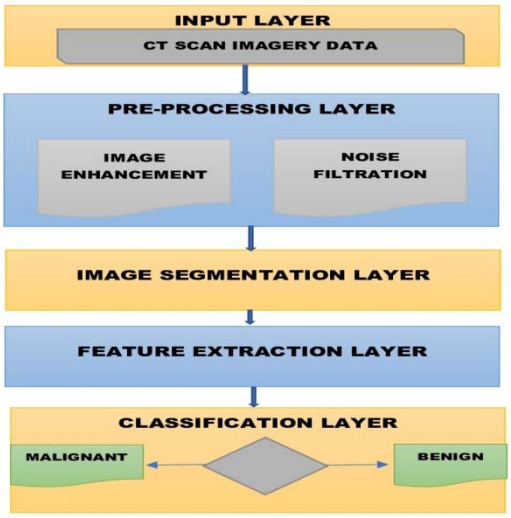

Fig 1. Proposed framework for lung cancer prediction

A. *Pre-Processing Layer*

Image has been collected from LIDC-LDRI. The original image was full of noise and for that first we have applied histogram equalization on the image to enhance the image and then on the equalized image median filter has been applied to remove the noise which was already present in the image after getting the noise free image we have applied some more noise in the image yield more clearer picture then again noise has been removed using median filter. Generally median filter is non linear digital filtering technique and it is also used as smoothing of images as it don't blur the edges completely as compare to other filtration technique like Gaussian filter or average filter.

B. *Segmentation Layer*

Image segmentation is a method of partitioning the image into various parts. After pre-processing the image on the pre-processed image segmentation is applied to acquire the information from the image. For image segmentation first we have applied edge detection technique through edge detection we can segment the boundary of the image for edge detection prewitt operator has been used, on that operator threshold has been applied so that after edge detection the intensity value which is less than threshold is removed and the intensity value which is higher than or equal to threshold will consider for further segmentation after getting the segmented image by edge detection we will apply watershed segmentation on the output image. Watershed segmentation takes the concept topographical landscape with ridge and valley which is defined by a gray level with respective pixel or gradient magnitude. There exist various ways to segment using watershed segmentation here we have used watershed segmentation using gradient. The gradient magnitude is used to preprocess the gray scale image; it has high pixel value along the object edge and low pixel value in another left region. And through this we can get the final segmented image through which we can extract features.

C. *Feature Extraction Layer*

The output generated by segmentation is used for feature extraction. By doing feature extraction we have extracted two types of feature one is region based another is texture based region based we have extracted feature like area in context to image means pixel of the image, perimeter in context image mean vector containing the distance around the boundary of each region in the image, centroid means the centre of mass of the region and it is in 1 X 2 vector form, image and based on texture we have extracted feature like mean is used to find average intensity, standard deviation is used to measure average contrast, smoothness used to measure relative smoothness of the intensity in the region, entropy is used to measure randomness using statistical approach of texture based.

D. *Classification Layer*

After feature extraction we will apply classification technique on both the feature to compare at which feature extraction which machine learning algorithm is giving more accuracy. Machine learning algorithm which has been used is support vector machine, artificial neural network and Random forest. After applying classification technique, it can be predicted that the tumour is cancerous or not and at which feature we are getting more accurate prediction.
Proposed Algorithm can be viewed as follows:

Input: Image Data (ID)
Output: Classification as benign or malignant

***Step 1**: Input the image data (ID)*
***Step 2**: Pre-Process the image*
***Step 2.1**: If the image is noise free*
Go to step 3
Else
Go to step 2.1
***Step 2.2**: Apply image Enhancement Method*
***Step 2.3**: Apply filter to enhanced image to reduce noise*
***Step 3**: Segment the image*
***Step 3.1**: Segment the boundary of the output image generated at step 2.3 using Edge Detection*
***Step 3.2**: After edge detection segment apply watershed gradient segmentation.*
***Step 4**: Feature Extraction*
***Step 4.1**: Region based feature are extracted like area, perimeter, centroid.*
***Step 4.2**: Statistical based feature are extracted like mean, standard deviation, smoothness.*
***Step 5**: Apply classification algorithm for training and prediction of tumour as benign or malignant.*
***Step 6**: Evaluate the parameter like accuracy, precision, Recall.*
***Step 7**: End*

V. PERFORMANCE EVALUATION

Model evaluation metrics are used to evaluate the performance of the model. The choice of metrics depends on the machine learning task. Types of model evaluation parameter are:

A. Confusion Matrix

Confusion matrix gives a detail description of classification or misclassification in a form of matrix. It consists of true positive (correctly predict the positive class), true negative (correctly predict the negative class), false positive (incorrectly predict the positive class), false negative (incorrectly predict the negative class).

B. Clasification Accuracy

It is used to measure the performance of our prediction. It can be measure by correct prediction by overall prediction made.

$$P_{acc} = \frac{T_{neg} + T_{pos}}{T_{pos} + F_{pos} + F_{neg} + T_{neg}}$$

C. Recall

It measures the proportion of actual positive that are correctly identified.

$$P_r = \frac{T_{pos}}{T_{pos} + F_{neg}}$$

D. Precision

It measure the proposition of positive identification is actually correct.

$$P_{prec} = \frac{T_{pos}}{T_{pos} + F_{pos}}$$

E. F1 score

F1 score is the average of both precision and recall.

$$P_{F1} = 2 * \frac{Recall * Precision}{Recall + Precision}$$

VI. RESULT AND DISCUSSION

In the proposed model for classification of tumour begin malignant or benign the machine learning algorithm used is artificial neural network, Random forest and Support vector machine. In both the feature that is region based and texture based artificial neural network is giving more accuracy. And comparing the accuracy with the proposed model, then it can be seen that accuracy has been increased whereas recall was less. For digital image processing was implemented in matlab R2017a and for classification using machine learning was implemented in jupyter notebook. A comparison between both the features is shown below

TABLE I. REGION BASED FEATURE

	Accuracy	Precision	Recall	F1 score
Random Forest	79%	100%	50%	67%
SVM	86%	100%	67%	80%
ANN	92%	100%	69%	81%

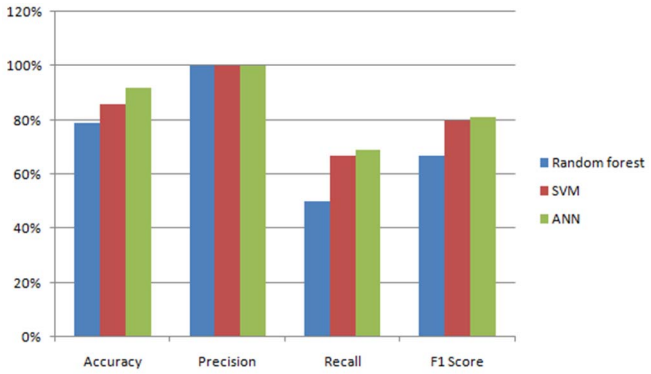

Fig 2. Performance measure based on region based extraction

TABLE II. TEXTURE BASED FEATURE

	Accuracy	Precision	Recall	F1 Score
Random Forest	70%	89%	47%	62%
SVM	80%	90%	57%	69%
ANN	96%	100%	69%	81%

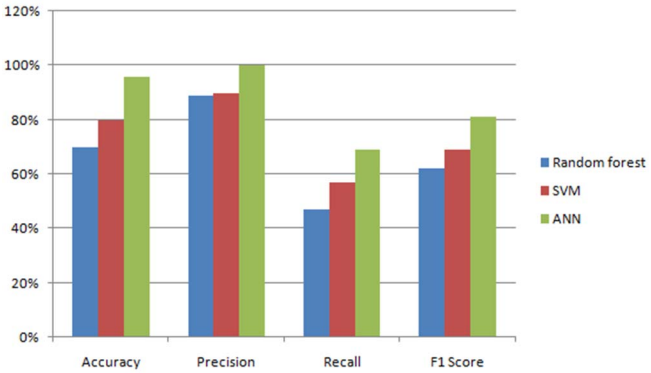

Fig 3. Performance measure based on texture based extraction

VII. CONLUSION AND FUTURE SCOPE

The proposed model shows the overview of prediction of lung cancer at an early stage. After prediction of the tumour begins malignant or benign, we generate a confusion matrix for each machine learning technique and based on the confusion matrix we calculate accuracy, Recall, precision and F1 score.

From the result we can say that our proposed model can distinguish between benign and malignant, and it can be seen that artificial neural network is providing more accuracy in both texture and region based, as well as from

the recall value we can say that it has correctly indentified maximum number of malignant tumour

In near future deep learning shall outperform machine learning in the field of image classification, object recognition and feature extraction. CNN networks are well-known for its features in providing accuracy with higher number of hidden layers in it.

REFERENCES

[1] Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." *International Journal of Computer Science and Information Technologies* 4.1 (2013): 39-45.

[2] Zhang, Junjie, et al. "Pulmonary nodule detection in medical images: a survey." *Biomedical Signal Processing and Control* 43 (2018): 138-147.

[3] Fenwa, Olusayo D., Funmilola A. Ajala, and A. Adigun. "Classification of cancer of the lungs using SVM and ANN." *Int. J. Comput. Technol.* 15.1 (2016): 6418-6426.

[4] Daoud, Maisa, and Michael Mayo. "A survey of neural network-based cancer prediction models from microarray data." *Artificial intelligence in medicine* (2019).

[5] Palani, D., and K. Venkatalakshmi. "An IoT based predictive modelling for predicting lung cancer using fuzzy cluster based segmentation and classification." *Journal of medical systems* 43.2 (2019): 21.

[6] Lynch, Chip M., et al. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." *International journal of medical informatics* 108 (2017): 1-8.

[7] Öztürk, Şaban, and Bayram Akdemir. "Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA." *Procedia computer science* 132 (2018): 40-46.

[8] Jin, Xin-Yu, Yu-Chen Zhang, and Qi-Liang Jin. "Pulmonary nodule detection based on CT images using convolution neural network." *2016 9th International symposium on computational intelligence and design (ISCID)*. Vol. 1. IEEE, 2016.

[9] Sumathipala, Yohan, et al. "Machine learning to predict lung nodule biopsy method using CT image features: A pilot study." *Computerized Medical Imaging and Graphics* 71 (2019): 1-8.