**An Efficient Approach for Detection of Lung Cancer through Image Processing**

Lavanya N Gautham Naik Apoorva Zaheen Ayesha Sharath Kumar

SMVITM, Bantakal

**Abstract – Lung cancer is one of the most common diseases in the world. It is very difficult to detect lung cancer at early stages. Statistics indicate that lung cancer, throughout the world, is the one that attacks the greatest number of people. Early detection of lung cancer is very important for successful treatment. In recent years the image processing mechanisms are used widely in several medical areas for improving early detection and treatment stages, in which the time factor is very important to discover the disease in the patient as fast as possible, especially in lung cancer. Along with image processing, we can also use machine learning techniques for detection of lung cancer. Various machine learning algorithms are considered and the machine is trained to detect the processed CT scan images. The efficient algorithm is chosen. In this paper we describe the image processing of the CT Scan images and the machine learning techniques applied to obtain an efficient method for early and accurate detection of lung cancer.**

**I Introduction**

Cancer is defined as a disease that highlights abnormal cell growth followed by unpredictable cell reproduction. Cancer is a dangerous and serious health problem in the whole world. Normal cell physiology constitutes the cell growth and cell division. If the cell growth and the cell division are at a rapid rate then it may lead to cancer. There are many types of cancers affecting various parts of the body starting from bone to blood. The mortality rate of lung cancer is the highest among all other types of cancer. It is one of the most serious and dangerous cancers in the world with the smallest survival rate after the diagnosis, with a gradual increase in the number of deaths every year. It is not easy to detect lung cancer at early stages. An estimated study says that nearly 85% lung cancer in men and 75% in women are caused due to cigarette smoking[1]. Also lung cancer can be caused to passive smoking. . People do have a higher chance of survival if the cancer can be detected in the early stages and they can be treated well.

Detection of lung cancer can be done using various procedures. But the radiologists cannot accurately diagnose lung cancer when looking at a CT scan. Interpretation of medical images is time consuming. As digital technologies are developing day by day they can be incorporated for the processing of these medical images. The implementation of image processing techniques and machine learning can make this process much more efficient. A Computerized Tomography (CT) scan is a combination of a set or a series of X-ray images. For detection and diagnosis of the lung cancer, CT scans are said to be more effective than plain chest X-rays. Unlike traditional X rays, which only highlight dense body parts, such as bones, CT also provides detailed views of the body’s soft tissues, including blood vessels, muscle tissue, and organs, such as the lungs. So we use CT scan images instead of X-Ray images.

This paper mainly includes the image processing techniques used to process the CT scan images. After the image processing is done the processed image is trained using various machine learning algorithms and compared. The best algorithm is chosen based on the efficiency of working on the test data set.

**II Related work**

Sun et al. compared three algorithms of deep learning and a traditional CAD system to diagnose lung cancer nodules on CT scan images. They used LIDC and IDRI datasets for the diagnosis of lung cancer. Deep Belief Network (DBN), CNN and Stacked Denoising Auto Encoder (SDAE) were used as three algorithms of deep learning. The results of the accuracy rates demonstrated that CNN and DBN were superior compared to the SDAE and traditional CAD methods Furthermore, DBN achieved the highest accuracy rate of nodules classification (81.19%).Since large image datasets of lung cancers are rare and deep learning methods are novel in the diagnosis of diseases, there are few researches in the diagnosis of lung cancer.

Ciompi et al. applied Multi-scale CNN with multi-stream architecture as a deep learning method for the classification of lung cancer nodules on CT scan images. In order to characterize lung cancer nodules, Multicentric Italian Lung Detection (MID) and Danish Lung Cancer Screening (DLCS) datasets were used. Automatic nodules classification in six types was done without using any segmentation methods. All scales of CNN were combined in a fully connected layer of CNN. Results of proposed multi-scale CNN were compared with radiologists’ diagnosis and the average accuracy rate of CNN (69.6%) is close to the average accuracy rate of radiologists (72.9%). Moreover, the accuracy of CNN with three scales was compared with SVM-based pixel intensity of patches and SVM-based unsupervised learning of features. Results show that CNN with three scales achieved a higher accuracy rate (79.5%) than the other two SVM based methods.

Yutong Xie et al. says that The accurate identification of malignant lung nodules on chest CT is critical for the early detection of lung cancer, which also offers patients the best chance of cure.The LIDC-IDRI database in the Cancer Imaging Archive (TCIA) contains 1018 clinical chest CT scans with lung nodules obtained from seven institutions. There is an associated XML file that details the locations of nodules on each 512×512 slice. They proposed MV-KBC algorithm consists of four major steps: (1) extracting 2D nodule slices from nine views of planes, (2) extracting the OA, HVV and HS patches on 2D nodule slices, (3) constructing nine KBC submodels and training each of them using the patches extracted on each view of planes, and constructing and training the MV-KBC model for lung nodule classification. The nodule appearance on nine view planes and the nodule heterogeneity and by applying an adaptive weighting scheme so that their model can be trained in an end-to-end manner. The results show that our model is more accurate than current state-of-the-art approaches on the LIDC-IDRI dataset.

Gavrielides et al. presented a technique based on an adaptive filter to estimate the size of the nodules and investigated which were the inter related factors that affect the accuracy in the measurement of pulmonary nodules. The main contribution of this paper is to present the main sources of error found in the measurement of pulmonary nodule, which may result in the appearance of new techniques. However, this research was restricted to solid nodules.

Peter Lu et al. they used the patient lung CT scan dataset from Kaggle’s Data Science Bowl 2017. The pixel values in each image to Hounsfield units (HU) and The segmentation obtained from thresholding has a lot of noise- many voxels that were part of lung tissue, especially voxels at the edge of the lung, tended to fall outside the range of lung tissue radio density due to CT scan noise. U-Net is a 2D CNN architecture that is popular for biomedical image segmentation. They designed a stripped-down version of the U-Net to limit memory expense.Once the U-Net was trained on LUNA 16 data[11], it was run on 2D slices of Kaggle data and stacked the 2 D slices back to generate nodule candidates.

Vijay A. Gajdhane et.al suggested multiple image procsssing methods.This procedure involved three processes majorly, Pre-processing, Feature Extraction and Classification. In the first step median filter was used to remove the noise, and with the aid of gabor filter image is segmented and enhanced. The features extracted were eccentricity, perimeter and area. In the last step, for classification SVM is used. In6 a SVM based classifier were proposed the help of image classification technique based on features.

Song et al. compared the performance of Deep Neural Network (DNN), CNN and Stacked Auto Encoder (SAE) algorithms in the classification of CT scan images of LIDC-IDRI datasets. The results showed that the CNN algorithm surpassed the other two algorithms in the classification of lung CT scan images. CNN, DNN and SAE achieved 84.15%, 82.37% and 82.59% accuracy rate. CNN and SAE achieved the same sensitivity (83.96%) and DNN achieved 80.66% sensitivity.

**III Image processing**

Digital image processing refers processing of the image in digital form. Modern cameras may directly take the image in digital form but generally images are originated in optical form. They are captured by video cameras and digitalized. The digitalization process includes sampling, quantization. Then these images are processed by the five fundamental processes, at least any one of them, not necessarily all of them.

First we need to produce a digital image from a paper envelope. This can be done using either a CCD camera, or a scanner. This is the step taken before the major image processing task. The problem here is to perform some basic tasks in order to render the resulting image more suitable for the job to follow. In this case it may involve enhancing the contrast, removing noise, or identifying regions likely to contain the postcode. Segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. Image description is the process of extract features that result in some quantitative information of interest or features that are basic for differentiating one class of objects from another. Image recognition is the process of assign a label to an object based on the information provided by its descriptors. Image interpretation is the process of assign meaning to an ensemble of recognized objects.

This section tells about the image processing methods used to process the CT scan images. In includes various stages of processing and is conducted using matlab.

The first step in image processing is image acqusitition. The first stage starts with taking a collection of CT scan images from the Database. Images are stored in MATLAB and displayed as a gray scale image. The lung CT images having low noise when compared to scan image and MRI image. So we can take the CT images for detecting the lungs. The main advantage of the computer tomography image having better clarity, low noise and distortion.

Fig 1: CT scan image

The next step involves conversion into grey scale. The RGB image is converted into grey scale image because the RGB format is not supported in Matlab. These grey scale images are black and white or consists of different shades of grey. It is done using matlab. Sometimes these images can even contain noises such as salt pepper noise, white noises etc. These disturb the process of image processing. So using various denoising algorithms the noises are removed as well.

 Fig 2: Grey scale image

Next step is to convert into binary image. Each pixel is just blackor white. Since there are only two possible values for each pixel (0, 1), we only need one bitper pixel. This image is just black and white.

Fig 3: Binary image

Next step involves image enhancement. Here we have removed the small objects and border removal is also carried out. In border removal the extra borders are removed and only the lung images are retained.

Fig 4: Image enhancement

After this the region of interest in the lung image is obtained and canny edge algorithm is applied. The image obtained after this is masked using the matlab and then the part where cancer is affected is obtained.



  Fig 6: Masked image and cancer image

**IV Machine learning algorithms**

In this paper we go through various machine learning algorithms used and the final result gives the best and efficient algorithm after testing the testing set **.**

Artificial Neural Network

The term "Artificial neural network" refers to an artificial intelligence subfield modeled after the brain that is biologically inspired. An Artificial Neural Network is usually a computer network focused on the biological neural networks that construct the human brain structure. Compared like a human brain, neurons are intertwined, artificial neural networks do contain neurons linked in separate layers of the networks. Such neurons are called nodes. An Artificial Neural Network in the world of Artificial Intelligence where it tries to emulate the cortical network makes up a human brain so that machines have an opportunity to comprehend things and make human-like decisions. The artificial neural network is programmed to act like interconnected brain cells simply by programming computers.

Artificial Neural Network can best be represented as a weighted, directed graph where the nodes are formed by artificial neurons. The relation between the outputs of the neurons and the inputs of the neurons can be interpreted with weights as the directed edges.

The Artificial Neural Network receives the vector-shaped input signal from the external source in the form of a pattern and image. For each n number of inputs, these inputs are then mathematically assigned by the x(n) notations.



Fig 7: Artificial neural network

Each data is then compounded by its corresponding weights (these weights are the information that artificial neural networks use to solve a particular problem). Such weights usually reflect, in general terms, the frequency of the neuronal interconnection within the artificial neural network. In the computing unit, all weighted inputs are summed. When the weighted sum is equal to zero, then distortion is applied to make the output non-zero, or anything else to scale up to the device response. Bias has equal input and weight equals 1. The total weighted inputs here can be in the range from 0 to positive infinity. Here a certain maximum value is benchmarked to hold the output within the limits of the desired value, and the cumulative weighted inputs are passed through the activation function. The activation function refers to the set of transfer functions that the desired output is achieved. There is a different type of activation function, but essentially either linear or non-linear function sets. The Binary, Linear, and Tan hyperbolic sigmoidal activation functions are some of the widely used activation sets.

Decision Tree

Decision Tree is a supervised method of learning used for the classification and regression processes of data mining. It is a tree which helps us to take decisions. The decision tree provides a tree structure for classification or regression models. It divides a data set into smaller subsets, thus constantly improving the decision tree. The final tree is a tree with nodes in decision and nodes in the leaf. At least two branches have a Decision Node. The nodes on the leaf suggest a designation or judgment. In a tree that relates to the best predictor called the root node, we can't accomplish more split on leaf nodes-The top decision node Decision trees can handle both the numerical and categorical data.



Fig 8: Decision tree

A decision tree is a hierarchical tree structure which can be used by implementing a sequence of simple decision rules to split an extensive collection of records into smaller sets of the class. A decision tree model includes a set of rules for dividing a huge heterogeneous population into smaller, more homogeneous, or classes which are mutually exclusive. Class attributes can be any variables from nominal , ordinal, binary, and quantitative values, while classes must be a qualitative type, such as categorical, or ordinary, or binary. In short, a decision tree creates the given attribute data together with its class, a set of rules that can be used to identify the class. One rule after another is applied resulting in class hierarchy within a group. The hierarchy is known as the tree, and they call each segment a node. With each progressive division, the members from the subsequent sets are becoming increasingly similar.

Support Vector Machine

Support vector machines (SVMs) are efficient machine learning devices for sorting and predicting results. The problem of separating two classes is handled using a hyperplane that maximizes the class margin. The margin-lying data points are called support vectors. The SVM algorithm seeks to find the hyperplane which creates the greatest margin for the two classes between the training points. It also penalizes the total distance of points on the wrong side of their margin when the two classes of data overlap. This allows to tolerating a limited number of false negatives near the margin. The other primary computational feature in SVM is the use of kernel functions and penalty parameter to translate nonlinear boundaries to linear boundaries in any higher-dimensional transformed space in the inputs' parameter space. The radial basis function is a popular choice in SVM applications.



Fig 9: SMV architecture

Here, the boundary demarcation between the classes red and blue (left panel) shows a predominantly continuous space for the class red with embedded blue pockets. The fitted SVM (right panel) model also creates a diagonally dominant pattern, though one in which the blue class is continuous. In both cases, the proportional fraction of blue over red space is very close.

Convolution Neural Network

The convolution neural Network is one of the key groups of neural networks to do image detection and pattern recognition. A few of the fields where convolution neural networks are commonly used are scene tagging, object detections, and facial recognition, etc.

CNN takes an image as an input that is classified and processed in a category like dog, cat, lion, tiger, etc. The computer sees an image as a pixel array, and depends on the image's resolution. It will look as h \* w \* d based on the image resolution, where h= height w= width and d= dimension. An RGB image, for instance, is 6 \* 6 \* 3 matrix array, and the grayscale image is 4 \* 4 \* 1 matrix array.



Fig 10: CNN architecture

Each input image in CNN will pass through a sequence of convolution layers together with pooling, fully connected layers, filters (also kernels). The Soft-max function then classifies an object with the probabilistic values 0 and 1.

**V Result and analysis**

After the image processing is done, if cancer is present it gives a message saying malignant and shows the cancer affected part. If the patient is not affected by cancer then it says benign.

The machine learning algorithms are used for classification and 4 different algorithms are used for training. After the algorithms are trained the accuracy is measured. depending on the accuracy and the performance the efficient algorithm is selected. This will be the future implementation of the project.

**References**

1. Comparative Analysis of Learning Algorithms for Lung Cancer Identification Syed Abbas Ali, Fatima Waheed , Wajahat Rehman , Sallar Khan, Marium Zia and Jawaria Sallar.
2. Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques Mr.Vijay A.Gajdhane , Prof. Deshpande L.M. Dept. of Electronics and Tele-communication Engineering, TPCT’s College of Engineering, Osmanabad, Maharashtra, India M.E. (Electronics) Dept. of Electronics and Tele-communication Engineering, TPCT’s College of Engineering, Osmanabad, Maharashtra, India .
3. Cancer Cells Detection Using Digital Image Processing Methods .Bhagyashri G. Patil PG Student, Department of Electronics Engineering, SSVPS BSD’s COE, Dhule, Maharashtra, India Prof. Sanjeev N. Jain Associate Professor, Department of Electronics Engineering, SSVPS BSD’s COE, Dhule, Maharashtra, India .
4. S. Kalaivani, Pramit Chatterjee, Shikhar Juyal, Rishi Gupta, “Lung Cancer Detection Using Digital Image Processing and Artificial Neural Network”, International Conference on Electronics, Communication and Aerospace Technology, ICECA, 100-103, 2017
5. S,Sasikala, M.Bharathi, B.R.Sowmiya, “Lung Cancer Detection and Classification Using Deep CNN”, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2278-3075, Volume-8, Issue-2S, December 2018
6. Majid Murtaza Noor and Vinay Narwal, “Machine Learning Approaches in Cancer Detection and Diagnosis
7. H. C. Shin, H. R. Roth, M.Gao, and L. Lu, “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning”, IEEE Trans. Med. Imag., vol. 35, no. 5, pp. 1285-1298, May 2016
8. Kumar D, Wong A, Clausia D. A., “Lung nodule classification using deep features in CT images. In: Computer and Robot Vision (CRV), 2015 12th Conference on pp. 133-138. IEEE (2015)
9. Sukhjinder Kaur, “Comparative study review on lung cancer detection using neural network and clustering algorithm”, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 4, Issue 2, February 2015
10. Sun, W.; Zheng, B.; Qian, W. , Computer aided lung cancer diagnosis with deep learning algorithms.In Proceedings of the International Society for Optics and Photonics Conference, San Diego, CA, USA 24 March 2016.
11. Ciompi, F.; Chung, K.; Riel, S.J.V.; Setio, A.A.A.; Gerke, P.K.; Jacobs, C.; Scholten, E.T.; Prokop, C.S.;Wille, M.M.W.; Marchianò, A.; et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning., Sci. Rep.2017,7, 46479. [CrossRef] [PubMed]
12. Yutong Xie, Yong Xia,Member, IEEE, Jianpeng Zhang, Yang Song,Member, IEEE, Dagan Feng,Fellow,IEEE, Michael Fulham, and Weidong Cai,Member, IEEE,”Knowledge-based Collaborative Deep Learning for Benign-Malignant Lung Nodule Classification on Chest CT’, DOI 10.1109/TMI.2018.2876510, IEEE,Transactions on Medical Imaging
13. Peter Lu, Department of Computer Science,Stanford University,peterlu6@stanford.edu, Albert Chon,Department of Computer Science,Stanford University,achon@stanford.edu, Niranjan Balachandar, Department of Computer Science, Stanford University, niranja9@stanford.edu,”Deep Convolutional Neural Networks for Lung Cancer Detection”.
14. Gavrielides MA, Zeng R, Kinnard LM, Myers KJ, Petrick N: Information-theoretic approach for analyzing bias and variance in lung nodule size estimation with ct: A phantom study. Med Imaging IEEE Trans 2010,29 (10):1795–1807
15. Song, Q.Z.; Zhao, L.; Luo, X.K.; Dou, X.C. Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images.J. Healthc. Eng.2017,2017, 1–7. [CrossRef] [PubMed]
16. Mr. Vijay A. Ghajdane, Prof. Deshpande L. M. "Detection of Lung Cancer Stages on CT Scan Images by Using Various Image Processing Techniques", IOSR Engineering Journal of Computer Engineering(IOSR-JCE) , Volume 16, Issue 5, Ver. III (Sept-Oct 2014) PP 28-35